

Modern Acoustics and Signal Processing

Jens Blauert
Jonas Braasch *Editors*

The Technology of Binaural Understanding



 *Modern Acoustics and Signal Processing*

 Springer

Modern Acoustics and Signal Processing

Editor-in-Chief

William M. Hartmann, East Lansing, USA

Series Editors

Yoichi Ando, Kobe, Japan

Whitlow W. L. Au, Kane'ohe, USA

Arthur B. Baggeroer, Cambridge, USA

Christopher R. Fuller, Blacksburg, USA

William A. Kuperman, La Jolla, USA

Joanne L. Miller, Boston, USA

Alexandra I. Tolstoy, McLean, USA

More information about this series at <http://www.springer.com/series/3754>

The ASA Press

The ASA Press imprint represents a collaboration between the Acoustical Society of America and Springer dedicated to encouraging the publication of important new books in acoustics. Published titles are intended to reflect the full range of research in acoustics. ASA Press books can include all types of books published by Springer and may appear in any appropriate Springer book series.

Editorial Board

Mark F. Hamilton (Chair), University of Texas at Austin

James Cottingham, Coe College

Diana Deutsch, University of California, San Diego

Timothy F. Duda, Woods Hole Oceanographic Institution

Robin Glosemeyer Petrone, Threshold Acoustics

William M. Hartmann (Ex Officio), Michigan State University

Darlene R. Ketten, Boston University

James F. Lynch (Ex Officio), Woods Hole Oceanographic Institution

Philip L. Marston, Washington State University

Arthur N. Popper (Ex Officio), University of Maryland

Martin Siderius, Portland State University

G. Christopher Stecker, Vanderbilt University School of Medicine


Ning Xiang, Rensselaer Polytechnic Institute



Jens Blauert · Jonas Braasch
Editors

The Technology of Binaural Understanding



 *Modern Acoustics and Signal Processing*

 Springer

Preface

Sound, devoid of *meaning*, would not matter to us. It is the information sound conveys that helps the brain to understand its environment. Sound and its underlying meaning are always associated with time and space. There is no sound without spatial properties, and the brain always organizes this information within a temporal–spatial framework. This book is devoted to understanding the importance of meaning for spatial and related further aspects of hearing, including cross-modal inference.

People, when exposed to acoustic stimuli, do not react directly to what they hear but rather to what they hear means to them.

This semiotic maxim may not always apply, for instance, when the reactions are reflexive. But, where it does apply, it poses a major challenge to the builders of models of the auditory system. Take, for example, an auditory model that is meant to be implemented on a robotic agent for autonomous search-&-rescue actions. Or think of a system that can perform judgments on the sound quality of multimedia-reproduction systems. It becomes immediately clear that such a system needs

- Cognitive capabilities, including substantial inherent knowledge
- The ability to integrate information across different sensory modalities

To realize these functions, the auditory system provides a pair of sensory organs, the two ears, and the means to perform adequate preprocessing of the signals provided by the ears. This is realized in the subcortical parts of the auditory system. In the title of a prior book,¹ the term *Binaural Listening* is used to indicate a focus on sub-cortical functions. Psychoacoustics and auditory signal processing contribute substantially to this area.

¹*The Technology of Binaural Listening*, J. Blauert (ed.), Springer and ASA Press, 2013.

The preprocessed signals are then forwarded to the cortical parts of the auditory system where, among other things, recognition, classification, localization, scene analysis, assignment of meaning, quality assessment, and action planning take place. Also, information from different sensory modalities is integrated at this level. Between sub-cortical and cortical regions of the auditory system, numerous feedback loops exist that ultimately support the high complexity and plasticity of the auditory system.

The current book concentrates on these cognitive functions. Instead of processing signals, processing symbols is now the predominant modeling task. Substantial contributions to the field draw upon the knowledge acquired by cognitive psychology. The keyword *Binaural Understanding* in the book title characterizes this shift.

Both books, *The Technology of Binaural Listening* and the current one, have been stimulated and supported by AABBA, an open research group devoted to the development and application of models of binaural hearing.²

The current book is dedicated to technologies that help explain, facilitate, apply, and support various aspects of binaural understanding. It is organized into five parts, each containing three to six chapters in order to provide a comprehensive overview of this emerging area. Each chapter was thoroughly reviewed by at least two anonymous, external experts.

The first part deals with the psychophysical and physiological effects of *Forming and Interpreting Aural Objects* as well as the underlying models. The fundamental concepts of reflexive and reflective auditory feedback are introduced. Mechanisms of binaural attention and attention switching are covered—as well as how auditory *Gestalt* rules facilitate binaural understanding. A general blackboard architecture is introduced as an example of how machines can learn to form and interpret aural objects to simulate human cognitive listening.

The second part, *Configuring and Understanding Aural Space*, focuses on the human understanding of complex three-dimensional environments—covering the psychological and biological fundamentals of auditory space formation. This part further addresses the human mechanisms used to process information and interact in complex reverberant environments, such as concert halls and forests, and additionally examines how the auditory system can learn to understand and adapt to these environments.

The third part is dedicated to *Processing Cross-Modal Inference* and highlights the fundamental human mechanisms used to integrate auditory cues with cues from other modalities to localize and form perceptual objects. This part also provides a general framework for understanding how complex multimodal scenes can be simulated and rendered.

²https://www.kfs.oeaw.ac.at/index.php?option=com_content&view=article&id=1072&Itemid=920&lang=de [last access August 30, 2019].

The fourth part, *Evaluating Aural-scene Quality and Speech Understanding*, focuses on the object-forming aspects of binaural listening and understanding. It addresses cognitive mechanisms involved in both the understanding of speech and the processing of nonverbal information such as Sound Quality and Quality-of-Experience. The aesthetic judgment of rooms is also discussed in this context. Models that simulate underlying human processes and performance are covered in addition to techniques for rendering virtual environments that can then be used to test these models.

The fifth part deals with the *Application of Cognitive Mechanisms to Audio Technology*. It highlights how cognitive mechanisms can be utilized to create spatial auditory illusions using binaural and other 3D-audio technologies. Further, it covers how cognitive binaural technologies can be applied to improve human performance in auditory displays and to develop new auditory technologies for interactive robots. The book concludes with the application of cognitive binaural technologies to the next generation of hearing aids.

Bochum, Germany
Troy, USA

Jens Blauert
Jonas Braasch

Contents

Forming and Interpreting Aural Objects: Effects and Models	
Reflexive and Reflective Auditory Feedback	3
Jens Blauert and Guy J. Brown	
Auditory Gestalt Rules and Their Application	33
Sarinah Sutojo, Joachim Thiemann, Armin Kohlrausch and Steven van de Par	
Selective Binaural Attention and Attention Switching	61
Janina Fels, Josefa Oberem and Iring Koch	
Blackboard Systems for Cognitive Audition	91
Christopher Schymura and Dorothea Kolossa	
Configuring and Understanding Aural-Space	
Formation of Three-Dimensional Auditory Space	115
Piotr Majdak, Robert Baumgartner and Claudia Jenny	
Biological Aspects of Perceptual Space Formation	151
Michael Pecka, Christian Leibold and Benedikt Grothe	
Auditory Spatial Impression in Concert Halls	173
Tapio Lokki and Jukka Pätynen	
Auditory Room Learning and Adaptation to Sound Reflections	203
Bernhard U. Seeber and Samuel Clapp	
Room Effect on Musicians' Performance	223
Malte Kob, Sebastia V. Amengual Garí and Zora Schärer Kalkandjiev	
Binaural Modeling from an Evolving-Habitat Perspective	251
Jonas Braasch	

Processing Cross-Modal Inference

Psychophysical Models of Sound Localisation with Audiovisual Interactions	289
--	-----

Catarina Mendonça

Cross-Modal and Cognitive Processes in Sound Localization	315
--	-----

M. Torben Pastore, Yi Zhou and William A. Yost

Spatial Soundscape Superposition and Multimodal Interaction	351
--	-----

Michael Cohen and William L. Martens

Evaluating Aural-Scene Quality and Speech Understanding

Binaural Evaluation of Sound Quality and Quality of Experience	393
---	-----

Alexander Raake and Hagen Wierstorf

The Language of Rooms: From Perception to Cognition to Aesthetic Judgment	435
--	-----

Stefan Weinzierl, Steffen Lepa and Martin Thiering

Modeling the Aesthetics of Audio-Scene Reproduction	455
--	-----

John Mourjopoulos

A Virtual Testbed for Binaural Agents	491
--	-----

Jens Blauert

Binaural Technology for Machine Speech Recognition and Understanding	511
---	-----

Richard M. Stern and Anjali Menon

Modeling Binaural Speech Understanding in Complex Situations	547
---	-----

Mathieu Lavandier and Virginia Best

Applying Cognitive Mechanisms to Audio Technology

Creating Auditory Illusions with Spatial-Audio Technologies	581
--	-----

Rozenn Nicol

Creating Auditory Illusions with Binaural Technology	623
---	-----

Karlheinz Brandenburg, Florian Klein, Annika Neidhardt, Ulrike Sloma and Stephan Werner

Toward Cognitive Usage of Binaural Displays	665
--	-----

Yôiti Suzuki, Akio Honda, Yukio Iwaya, Makoto Ohuchi and Shuichi Sakamoto

Audition as a Trigger of Head Movements	697
--	-----

Benjamin Cohen-Lhyver, Sylvain Argentièri and Bruno Gas

Intelligent Hearing Instruments—Trends and Challenges	733
Eleftheria Georganti, Gilles Courtois, Peter Derleth and Stefan Launer	
Scene-Aware Dynamic-Range Compression in Hearing Aids	763
Tobias May, Borys Kowalewski and Torsten Dau	
Index	801

Forming and Interpreting Aural Objects: Effects and Models

Reflexive and Reflective Auditory Feedback



Jens Blauert and Guy J. Brown

Abstract Current models of binaural hearing go beyond bottom-up-driven processing and, instead, use a hybrid approach by including top-down, hypothesis-driven algorithms. Such hybrid models first identify and characterize auditory objects. Out of these objects, the model infers an auditory scene, from which it can extrapolate understanding, form judgments, and initiate actions. For example, when embedded in a mobile robot, a binaural hearing system can provide the information needed to carry out search-and-rescue tasks. Further, such systems are able to make judgments, for instance, on the quality of experience in spaces for musical performances. As with humans, such actions and judgments are based on sets of references built from perceptual structures, inherent and acquired knowledge, and the intellectual capabilities of the systems—in other words, on the “brains” of the model systems and the knowledge contained in them. To achieve these goals, adequate feedback loops must to be considered, evaluated, and implemented within technological models of auditory systems. In this chapter, a number of such feedback loops are described and discussed that have already been implemented and evaluated. A distinction is made between reflexive and reflective feedback mechanisms, the latter, including cognitive activities.

1 Introduction

The structure and function of the human auditory system have long been the subject of intensive scientific research.¹ Yet, most investigations have looked at it in isolation, disregarding that it is an embedded component of a much larger and more

¹For overviews see, for example, Moore (1995, 1989), Yost (2007), Plack (2010), Celesia and Hickok (2015) and Fay and Popper (1992–2017).

J. Blauert (✉)
Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
e-mail: jens.blauert@rub.de

G. J. Brown
Department of Computer Science, University of Sheffield, Sheffield, UK

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_1

complex system, namely, the human body. Further, when considering the function of the auditory system, there has been a tendency to focus on *afferent* (ascending) neural pathways. This led to concepts that favored “bottom-up-processing” models of the auditory system. *Efferent* (descending) pathways have often been neglected in accounts of auditory function (He and Yu 2009).

In fact, it has become evident that efferent pathways are as common in the auditory system as afferent ones (Schofield 2009). At some stages of the system, for example in the cochlea, the efferent fibers considerably outnumber the afferent ones (Shamma 2013). Various efferent pathways have been identified, connecting all stages of the system from the brain down to the cochlea and the middle ear. Some particularly relevant ones are schematically plotted in Fig. 1. Accordingly, top-down processes must be included in hypotheses regarding the function of the auditory system. Further, the coexistence of bottom-up and top-down processing is a strong indication of the existence of feedback loops.² In contrast to linear time-invariant systems, the auditory system thus turns out to be time-variant and nonlinear. In light of the presence of feedback processes, one could thus actually apply the somewhat old-fashioned term *cybernetic system*³ to it.

Feedback in biological systems is manifested in myriad, albeit occasionally incomprehensible, ways. Yet, from a technological point of view, those functions of feedback are of particular interest which may serve a purpose in given application scenarios. From a systematic point of view, it is useful to group the feedback paths roughly into two categories, namely, *reflexive* and *reflective* feedback as introduced in Two!Ears (2014), pp. 46–47 as follows.

- “*Reflexive feedback* is triggered by primitive perceptual cues without cognitive processing, that is, in a reflexive way. This kind of feedback reacts within a comparatively short time and will decay when the cues that trigger it have stopped. This happens regardless of a specific task. Also, reflexive feedback does not need, and most likely cannot even receive training. It has been supposed that the cues that trigger reflexive feedback cannot be ignored”
- “*Reflective feedback* requires reflection in the mind, that is, it puts a cognitive load on the modeling system. The feedback will be initiated with the goal of supporting a given task. Due to the processes involved (e.g., attention), this kind of feedback needs more time to react. Also, the cues that initiate the feedback may be memorized and remain effective almost indefinitely, since reflective feedback may not only react to sensory objects (such as auditory, visual, tactile, olfactory or gustatory ones) but also to emotions (feelings), and thoughts (ideas, concepts, notions)”

²That is, of loops where the output of a system or system element is routed back to its input with the effect of modifying the output.

³Cybernetics: The science of communication and control among biological organisms and machines.

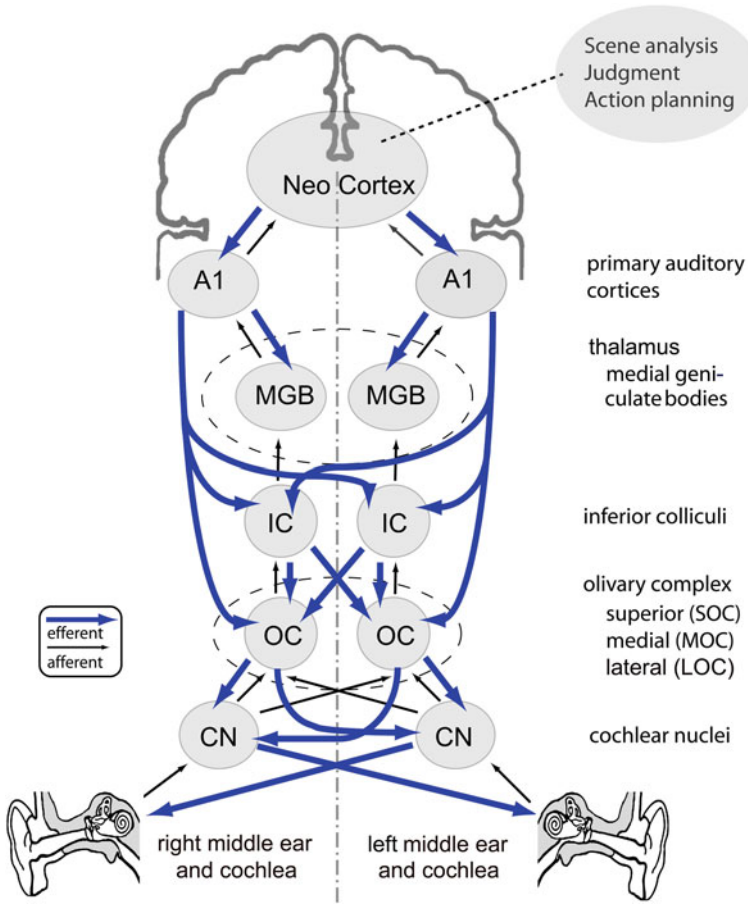


Fig. 1 Schematic of the auditory system with major afferent and efferent links

2 Modeling the Auditory System

This section explains how the research group “*Aural Assessment By means of Binaural Algorithms*” (AABBA)⁴ and the TWO!EARS project⁵ have developed a mobile robot with binaural listening capabilities by applying an integrated model of the auditory system. The robot has the ability to perceive, interpret, understand, and evaluate its environment through a combination of signal-driven and hypothesis-driven algorithms, and does so through exploratory movements. In order to exploit biological processes for technological applications, it is a common approach to model them by

⁴AABBA: <https://www.kfs.oeaw.ac.at/aabba> [last accessed, August 18, 2019], see also Blauert et al. (2010).

⁵TWO!EARS: <http://www.twoears.eu> [last accessed, August 18, 2019], see also Blauert (2017).

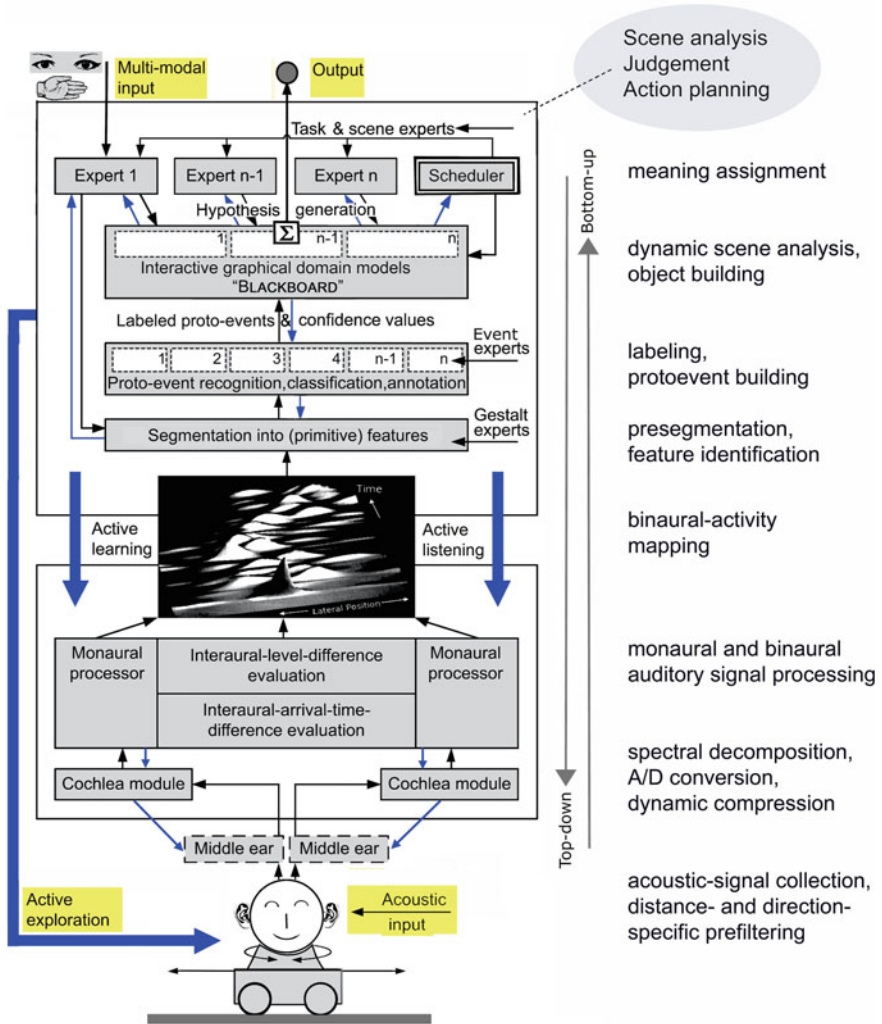


Fig. 2 Block diagram for a model of the auditory system including feedback (Raake and Blauert 2013)

technological means, that is, in terms of hardware and software. Figure 2 shows the block diagram of the model. The project considered the auditory system as part of an intelligent, multimodal artificial agent that actively explores the world. In the course of this process, the agent interprets what it perceives, collects knowledge, and develops its own concepts accordingly. Such an approach required, among other features, means for the exploration of the environment by situation-specific adaptations and cognition-controlled head-and-torso movements. More details of the model structure are reported in the following.

2.1 Model Architecture

The model employs a head-and-torso simulator (HATS) with replicas of human pinnae and internal microphones. A motor is integrated into the collar of the HATS to allow for rotational head movements (Bustamante et al. 2016). When mounted on a cart, the unit is capable of motion with two further translational degrees of freedom. To include visual input, a stereoscopic pair of cameras is available to allow for simulation of human binocular vision. The microphone signals are passed through middle-ear modules to a signal-processing chain, the functionality of which includes monaural and binaural processing in terms of masking, compression, modulation-spectrum analysis, interaural time- and level-difference analysis, and interaural coherence analysis. This results in a multidimensional auditory representation (binaural-activity map) which serves as the basis for subsequent dynamic auditory-scene analyses.

The scene-analysis components are organized as follows (Raake and Blauert 2013). One or multiple parallel processors carry out a presegmentation of the multidimensional feature representation, identifying key features for object and event identification. Examples of features are onsets, modulation characteristics, harmonic components or interaural time- and level-differences that are temporally collocated across different spectral bands and may, thus, be associated with particular objects. This stage can be seen as the lowest level at which primitive schemata are extracted, applying rule sets as those of *Gestalt* formation (Bregman 1990). For examples of Gestalt heuristics in the context of engineering applications, see Jekosch (2005) and Sutojo et al. (2020), this volume. Subsequently, the primitive features are interpreted in terms of auditory events.⁶

The next stages represent hypothesis generation, adaptation, and verification. To this end, symbolic information rendered by the lower stages is forwarded to a so-called “Blackboard system” Erman et al. (1980), Corkill (1991).⁷ This information is then accessed by a multi-expert system—see Fig. 2. Each expert analyzes the information in the Blackboard based on its respective expertise, identifying whether this information corresponds to known information. Areas of expertise examples include acoustics, psychoacoustics, object identification, psychology, spatial hearing, cross-modal integration, proprioception, speech communication, music, and sound quality.

The expert system is based on task- and context-specific information. It creates the symbolic information that describes the auditory scene. Using top-down feedback mechanisms, the experts can exert modifications in the symbolic representation at Blackboard-level or modifications at lower levels, for example, in terms of the respective feature-selection process (e.g., reflecting auditory attention), or the acoustic front-end (e.g., for turning the head). Accordingly, the Blackboard and the

⁶The events are referred to as *protoevents* here, reflecting their still existing statistical uncertainty at this processing stage.

⁷For more details on Blackboard systems see Schymura and Kolossa (2020), this volume.

group of experts represent the world-knowledge of the modeling system, as well as respective decision-units that can initiate actions.

It should be mentioned in passing that such a model system is not only useful as a basis for technological applications but may also serve to predict potential effects of presumed biological feedback loops. The Blackboard-system architecture has been selected on purpose for this research-based model, as it provides more transparency than data-driven end-to-end solutions such as Deep Neural Networks (DNNs)—see Sect. 3.3.

2.2 Feasible Feedback Loops

The two most relevant purposes of comprehensive and versatile models of the auditory system are, (a), to improve our understanding of how biological auditory system function, and, (b), to explore application opportunities in technological systems. In this context, two questions appear. Firstly, for which purposes would feedback loops be useful? Secondly, where in the model-system structure could they be implemented? Regarding the first question, feedback may be employed to increase the precision of, and reduce ambiguities in perceiving, reasoning, and acting. For both questions, the following systematic analysis provides general conceptions.

- (a) Problems at the signal's level: *Variances too high*
 ⇒ Approach: Follow causal links in the Blackboard system to identify relevant sources of uncertainty. Attempt to minimize variance in these observables.
- (b) Problems at the symbolic level: *Logical inconsistencies*
 ⇒ Approach: Identify causal links leading to competing hypotheses. According to the Blackboard-system structure, identify additional input necessary for conflict resolution.

Table 1 provides possible inputs to feedback loops, along with expected improvements. Further, expectations are formulated regarding what might be achievable by suitable feedback loops when using these inputs—see also Blauert and Obermayer (2012) and Blauert et al. (2014).

Table 2 provides a listing of possible entry ports for feedback loops and specifies what could possibly be controlled by proper use of these ports. The entry ports have been selected in view of modeling efforts rather than as a description of feedback loops in biological systems.

The two tables are meant as road maps for further investigations in auditory-feedback modeling. Only a few of the feedback loops mentioned here have been implemented and tested in technological systems so far—see Sect. 3.

Table 1 Points of origin of feedback-control information for respective feedback-loop entry ports, and expected functional improvements

Source of feedback signals or symbolic feedback information	Expected functional improvements
– Binaural-processing stage	– Turning the acoustic sensors into an optimal position
– Visual cues from the robot’s cameras	– Advanced movements of the head-&-torso platform (active exploration)
– Cues from the Blackboard system	– Exploiting contents of the graphical models and the knowledge sources
– Modules operating on the olivary system SOC/MOC/LOC level	– Increasing the signal-to-noise ratio, increasing spectral and temporal selectivity
– Presegmentation stage and Blackboard system (graphical model, knowledge sources, scheduler)	– Paying attention to specific signal features to deliver specific additional information as required by the cognitive stage
– Binaural-activity-mapping stage	– Activation of specialized, task-specific signal-processing procedures, such as echo cancelling, dereverberation, precedence-effect preprocessing
– Presegmentation stage and blackboard system	– Re-evaluation (reconsideration) to solve ambiguities
– Visual cues from the robot’s cameras	– Optimal positioning of the head-&-torso platform (task-specific)
– Sensorimotor cues from the head-&-torso platform	– Improvement of object recognition, auditory grouping, aural stream segregation, aural scene analysis, attention focusing
– External knowledge sources – Optical and acoustical information from the microphones and cameras of the head-&-torso platform	– Improvement of scene understanding, assignment of meaning, quality judgements, attention focusing

3 Selected Feedback Loops

Feedback in the auditory system has become a progressive item of research interest since many facets of the system performance need the assumption of feedback to be understood. It can thus be anticipated that modeling of auditory feedback and technological application of the respective models will increase. The following sections present a selection of feedback loops which have been tackled for modeling so far.

3.1 Input-gain Control

The auditory system comprises various means for controlling the amplitude of the signals and, hence, the rate of neural spikes that it processes. Two of them are placed

Table 2 Examples of entry ports for feedback and possible actions induced

Entry ports for feedback	Possible actions induced
Monaural and binaural processing stages (OC-level modules)	<ul style="list-style-type: none"> – Adjustment of time-windows, time constants and spectral regions – Task-specific employment of additional processing steps, e.g., lateral and contralateral inhibition, precedence preprocessing, dereverberation
Binaural-activity-mapping stage (IC-level modules)	<ul style="list-style-type: none"> – Setting time constants for contralateral inhibition – Providing masks for dedicated analyses of binaural-activity maps – Focusing on specific spectral regions – Adjustment of operation points and dynamic ranges – Provision of non-auditory sensory data, e.g., from vision, proprioception, sensorimotor cues
Blackboard architecture (graphical model, knowledge sources, scheduler)	<ul style="list-style-type: none"> – Provision of external knowledge, e.g., salient features, object-building schemata, rule systems – Knowledge of the situational history, communicative intention of sound sources – Task-specific expert knowledge, internal references – Provision of non-auditory knowledge, e.g., from visual scene analyses

fairly peripherally, namely, the *Middle-Ear-Muscle Reflex*, also known as the acoustic reflex or stapedius reflex, and the *Medio-Olivocochlear Reflex*. Both are mainly based on reflexive feedback but also react reflectively to activities at higher stages of the auditory system. There is potential for technological application because input-gain control may help shift the point of operation of the auditory system into a region of high discrimination of amplitude differences.

Yet, before discussing signal-amplitude and/or spike-rate control in more detail, it should be recalled that the auditory system is not linear. The inner ears, for example, perform a running spectral decomposition of the incoming signals and convert the spectral components into a series of neural spikes (amplitude-to-rate-code conversion). Thereby a specific compression takes place. Figure 3 shows some of the effects of amplitude variation in this context. The shape and bandwidth of the ear-filter-transfer functions vary with the amplitude of the acoustic input signals. The signal-to-spike-rate conversion is compressive. An obvious indication of these and further nonlinear processes is the shape and spacing of the equal-loudness contours. They become flatter, and their mutual distance gets larger with increasing sound-pressure level, an effect that is most pronounced at low frequencies.

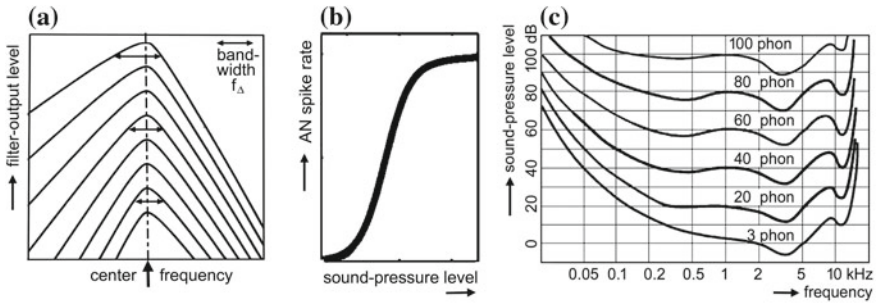


Fig. 3 **a** Shape and bandwidth of the ear-filter transfer functions are level dependent. **b** The neural-spike rate in the auditory nerve (AN) follows a compressive function with respect to the sound-pressure level at the eardrum. **c** Consequently, shape and spacing of the equal-loudness contours (isophones) are level dependent—compare ISO 226:2003 (ISO 2003)

3.1.1 The Middle-Ear-Muscle Reflex

The *Middle-Ear-Muscle Reflex* (MEMR) receives control signals from both the ipsi- and contralateral cochleae. The reflex arches are depicted in Fig. 4. They ascend through the Cochlear Nuclei (CN) up to the Olivary Complex (OC), and there mainly to the two Medial Olivary Complex (MOC) regions. Close to these regions, motoneurons emanate that project to the two middle-ear muscles, namely, *stapedius* and *tensor tympani*. Consequently, they can initiate the contraction of these. Due to such contractions, the stiffness of the ossicular chain increases, and so does the tension of the eardrum. This causes the middle-ear impedance to rise with the effect that more sound is reflected back into the ear canal instead of proceeding into the inner ear (cochlea). Consequently, the sensitivity of the auditory system declines temporarily. More information about anatomical and functional detail can be found in Møller (1962, 1974), Borg and Counter (1989), and Mukerji et al. (2010).

The MEMR can be triggered from each of the two ears and affects the ipsilateral as well as the contralateral middle ear. It is, therefore, used as a diagnostic indicator to, for example, discriminate hearing disorders in the middle ear and/or cochlea from those that reside further up in the auditory system—in the olivary complex, for instance. Activation of the MEMR can be detected by measuring variations of the eardrum impedance (Kung and Willcox 2007). In humans, it is predominantly the stapedius that is acoustically excitable, namely, by intense low-frequency sounds. As to the tensor tympani, there is some indication that it may, besides acoustically, also be triggered by crossmodal cues, such as visual ones (Mukerji et al. 2010; Djupesland 1964). It has further been shown that the MEMR can be activated from higher stages of the auditory system involving cognition, for example, when speaking in a noisy environment (de Andrade et al. 2011a, b). Some people can even activate their tensors tympani intentionally—a striking example of reflective feedback.

Activation of the MEMR results in an amplitude reduction of up to 10 dB, mainly in the low-frequency range. It has been suggested that the reflex plays the role of a hear-

ing protector in conditions which would otherwise constitute harmful noise levels. The reflex becomes active for sound-pressure levels from about 70 dB upward. Yet, one has to consider that the reaction-time amounts to roughly 25 ms to 150 ms—the higher the sound pressure level, the shorter the latency interval. Thus, for protection against harmful noise, it is necessary to predict its appearance before it actually starts. An interesting application is reported in Mercedes (2002). Lutman and Martin (1979) and Longtin and Derome (1989) have developed models for the MEMR which are useful for the specification of further technological applications.

3.1.2 The Medial-Olivocochlear Reflex

The *Medial-olivocochlear Reflex* (MOCR) controls the gain of the output signals of the inner ears (cochleae), namely, the rate of the neural spikes being transmitted from the cochlea to higher stages of the auditory system. This is performed by damping the movements of the basilar membrane via the outer hair cells. In addition, dendrites of afferent auditory-nerve fibers are affected (Guinan 1996; Guinan jr 2010)—see also Two!Ears (2014) pp. 75–66. The prominent reflex arch, shown in Fig. 4, is as follows—compare Brown et al. (2003). The output of each cochlea passes through the auditory nerve via the Cochlear Nucleus (CN) on to the Olivary Complex (OC). Within the OC these outputs go to left and right subsystems called Medial and Lateral Olivary Complexes (MOC and LOC). While the MOCs project to the outer hair cells in the cochlea, the LOCs do so to the inner ones. From the perspective of prospective technological application, the MOC path is the more relevant one.

The MOCR can be excited ipsi-, contra-, or unilaterally. The necessary sound-pressure level to activate it is lower than the one that is needed to excite the MEMR.

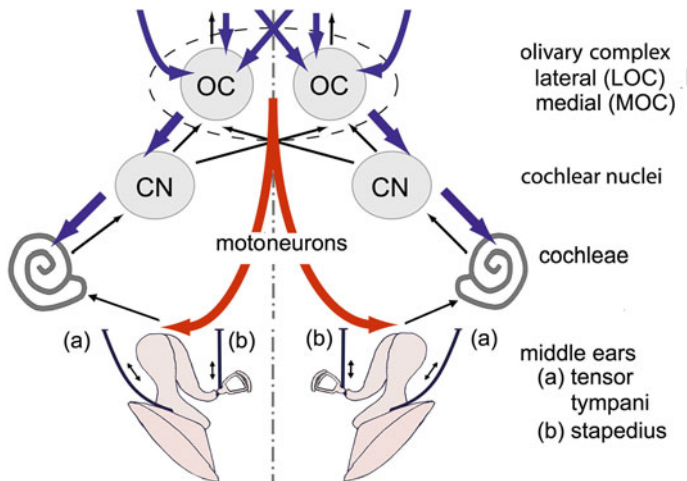


Fig. 4 Feedback loop of the middle-ear and the medio-olivocochlear reflexes

Contralateral excitation is more pronounced than ipsilateral excitation. In most studies, levels on the order of 40–50 dB have been shown to be sufficient to trigger the effect significantly. This guarantees that MOCR effects are not mixed up with MEMR effects experimentally. The action principle of the MOCR is, roughly speaking, as follows. The outer hair-cell activity is reduced, and thus, cochlear amplification is also diminished. The basilar membrane is, therefore, damped so that its velocity decreases. Consequently, the inner hair cells, which are, so to say, the acoustic sensors (microphones) in the cochlea, reduce their sensitivity. This shifts the point of operation for the acoustic inputs signals on the resulting sound-pressure-level-to-spike-rate function (compare Fig. 3a) to the section with a steeper slope. Consequently, a low-level noise and a desired signal with at least a slightly higher level improve the perceptual separation. It is widely understood that this increase of the perceptual signal-to-noise ratio is the key physiological function of the MOCR. As to the reaction times of the MOCR, there seems to be a fast one of 10–100 ms and a slow one of up to 100 s (Zhao and Dhar 2011). They obviously represent different reaction principles, but these are not yet convincingly identified.

A number of computer models of the MOCR have been proposed, for instance by Ferry and Meddis (2007) and Ghitza et al. (2007). They model the amount of reduction of the basilar-membrane velocity due to the MOCR by a network with two parallel branches, one of them is linear and time invariant, the other one is nonlinear and time-variant. Ferry and Meddis (2007) refer to this structure as a *Dual-Resonance Nonlinear Model* (DRNL). The nonlinearity is caused by a “broken-stick” input-output characteristic with a steep middle part and shallow legs at the upper and lower ends—see the box labelled **b**₃ in Fig. 5. Both branches receive the same acoustic input signals, and their two outputs are superimposed. The nonlinear branch includes a controllable attenuator. If attenuation is set to maximum (i.e., the lower branch is switched off), MOCR suppression is fully effective. When the attenuation is reduced, suppression becomes smaller, but at the same time nonlinear distortions of the output increase.

The MOCR provides reflexive feedback, but there is some indication that it can also be controlled by top-down information from higher stages of the auditory system,

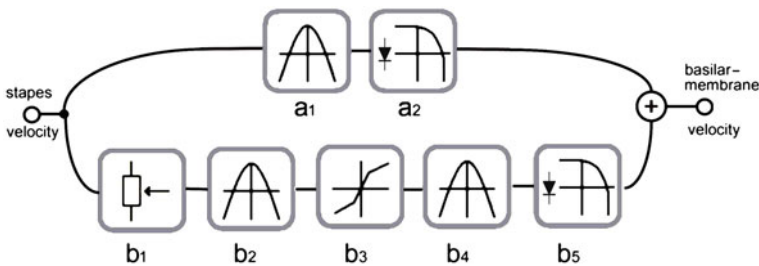


Fig. 5 Descending part of the Medio-Olivocochlear reflex arch—after Brown et al. (2010). Linear path **a**₁, **a**₂ with band-pass filter, and rectifier-&-low-pass filter. Nonlinear path **b**₁ . . . **b**₅ with attenuator, band-pass filter, compressor, band-pass filter, and rectifier-&-low-pass filter

that is, reflectively. Attention plays a role in this context—see, for example, Harkrider and Bowers (2009) and Smith et al. (2012).

With regard to technological applications, a specific electronic model of the reflex has recently been developed that improves speech perception in noise for cochlear implantees, that is, for hearing impaired persons that rely on electric rather than acoustic stimulation of their auditory nerves. Noise (or even undesired speech) is fed to the cochlear-implant processor at one ear, and the gain of the contralateral processor is then reduced by mimicking the MOCR reflex. The perceptibility of speech delivered to the contralateral ear is then improved, even if the noise and the speech sources are positioned in the same direction with respect to the listener (Lopez-Poveda et al. 2016, 2017). The MOCR has also been used as the basis for a hearing-aid algorithm that has been implemented by Meddis and colleagues on low-cost smartphone hardware (Jurgens et al. 2016).

3.2 *Simplified Cochlear Models for Technological Systems*

The output of the two inner ears (cochleae) represents the prominent input to the auditory nerves and, thus, to higher stages of the auditory system. As has been mentioned before, the cochleae decompose their acoustic input into neural-spike trains, organized into spectral bands (Sect. 3.1). The cochleae are the effectors, that is, the last element of the chain, for various feedback loops that originate at higher auditory processing stages. Further, there are feedback loops in the inner ears themselves, for instance, from the inner to the outer hair cells. Nevertheless, cochlea models, as used in technological applications, are usually quite simplified as compared to the biological complexity of the cochlea.

For technological purposes, a cochlea is usually modeled as a bank of band-pass *ear filters*. These are commonly realized as so-called *Gammatone filters*, that is filters with an impulse response consisting of a sinusoidal function with a gamma-distribution-shaped envelope (Holdsworth et al. 1988). Gammatone filters can be implemented efficiently and roughly mimic the impulse responses of biological ear-filters. They are mathematically defined by the expression

$$g(t) = at^{n-1}e^{-2\pi fbt} + \cos(2\pi ft + \phi), \quad (1)$$

where n is the filter order, f is the center frequency (in Hz), b is the bandwidth (in Hz), t is time (in seconds), ϕ is the phase of the carrier, and a is the amplitude.

Usually, the output signal of each spectral band is rectified and fed through a low-pass filter (cut-off frequency ≈ 600 Hz, moderate slope). This results in amplitude demodulation of the band-pass output signals. Thus, above the low-pass cut-off frequency the envelope of the signals will be extracted. So far, such a model system behaves linearly. For instance, the filter shape is not level dependent. Yet, for better approximation of the biological case, some models weight the output signals according to the sound-pressure-level-to-spike-rate function—see Fig. 3b. This

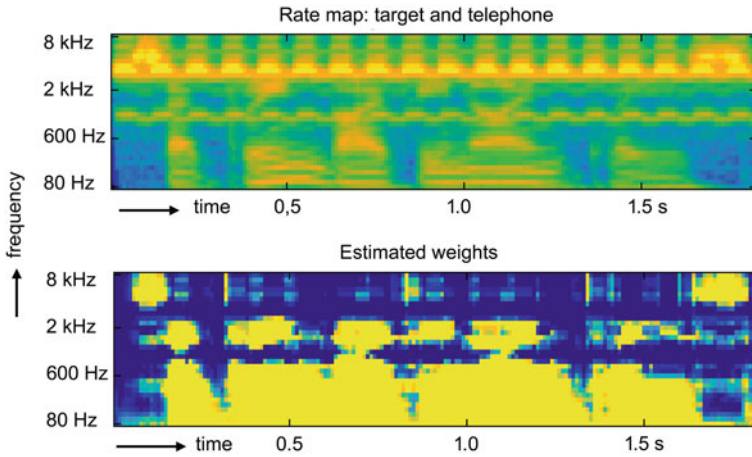


Fig. 6 Example: rate map and estimated weights for a 1.5-s target-speech sample interfered by telephone ringing. The **yellow regions** depict target dominance

causes compression and, thus, level-dependency. As its output, the model thus performs a frequency analysis similar to a running Fourier transform or weighted running Fourier transforms (magnitudes only) of the two acoustic cochlea-input signals. The results can be visualized as plots with the coordinates frequency and running time, and with the amplitudes or spike rates, respectively, coded in color or gray scales. Such plots are commonly termed *rate maps* in the field (Brown and Cooke 1994)—see upper panel of Fig. 6 as an example.

The fact that the sensitivity of biological cochleae is controlled from higher stages of the auditory system is usually not considered in technological cochlea models by modifying the pressure-level-to-spike-rate function. Rather, it is realized by assigning suitable weights to the output spike rates in the individual spectral band before further processing steps are taken.

3.3 Auditory-Object Localization Assisted by Head Movements

Perceptual objects are defined by their essential features, their position and spatial extent in the perceptual space, and their relation to other objects. Objects are spatially and temporally distinct. In other words, they exist at a certain time in a certain position. Localization is thus a basic requirement of object formation—refer to Blauert (1997) for fundamentals of auditory localization. Auditory localization was traditionally discussed as a static phenomenon, not considering that the ears are positioned on a head which is movable in six degrees of freedom. In fact, the exploitation of additional cues as collected by head movements improves the local-

ization capabilities considerably, for instance, in the course of exploring unknown environments. For example, head movements are utilized to solve directional ambiguities such as front-back confusion (Perrett and Noble 1997) and to put the head into a suitable position for segregation of desired signal components from undesired ones, such as noise, reverberation, and/or concurrent talkers (Braasch et al. 2011; Parks and Braasch 2013).

3.3.1 The “Turn-to-Reflex”

In the case that a sudden sound occurs in the vicinity of listeners, it is a common reaction that they turn their heads towards the direction of the sound source. This behavior is often called *Turn-to-reflex*. This action helps to put the sound-source into the visual field and thus helps to identify the reason for the sound. The concept of *Attention Reorientation* is noteworthy in this context (Corbetta et al. 2008)—compare also Cohen-L’hyver et al. (2020), this volume.

Interestingly, blind people usually do not turn the head in a frontal position to the source, but rather slightly to the side, namely, in the direction of best hearing.⁸ Whether this is still purely reflexive or also reflective is worthy of discussion; thus the term turn-to-“reflex” is slightly misleading. It appears that there are different neural streams involved, depending on the goal-driven or stimulus-driven attentional behavior of the person considered (Petersen and Posner 2012).

In any case, appropriate head turning requires fast and reliable auditory source localization, also in situations with concurrent sources. To model this behavior, different approaches have been implemented and tested. These algorithms typically try to determine the directions of sound-wave incidence on the listeners, although the directions are not necessarily spatially coincident with the directions in which the auditory objects are actually heard. Evidently, the following two are particularly effective.

(i) *Deep-Neural-Network-Based Localization*

Deep neural networks (DNNs) are neural networks with several hidden layers, which contain representations of different degrees of abstraction. They are a versatile tool for many classification tasks when working on large data sets (Goodfellow et al. 2016). In auditory localization, they have proven successful for the identification of the directions of multiple sound sources, such as concurrent speakers, even in noisy and reverberant environments.

An example is the system of Ma et al. (2015), see also Ma and Brown (2015, 2016), in which a DNN is used to learn the relationship between the source azimuth and binaural cues, namely, the cross-correlation function and interaural level differences of the signals arriving at the two ears. The DNN was trained using a multi-condition approach; spatially diffuse noise was added to the training signals at different signal-to-noise ratios in order to improve robustness to reverberation. The authors show that

⁸Personal observation (first author).

their system is able to accurately localize a target source in challenging conditions, that is where multiple talkers and room reverberation are present.

(ii) Null-Steering-Antenna-Based Localization

The null-steering-antenna approach is a method that is useful for foreground-background separation. It is, for instance, widely used for beamforming of antennas—see, for instance, Leng et al. (2008). In psychoacoustics the approach was introduced in modified form by Durlach (1960) as *equalization-cancellation* theory. The basic idea is explained here in its “cophase-and-subtract” version (Mi et al. 2017). Consider two microphones spaced apart from each other by roughly the distance of the two ears. If a sound-source is positioned perpendicular to this two-microphone array, the signal components stemming from this source and arriving at the two microphones will be identical. If now one microphone signal is subtracted from the other one, the result will be null. This directional null indicates the source position.

Localization Enhancement by “Sharpening the Ears”

The following two examples of feedback in sound-source localization exploit the fact that the pattern of acoustic events or characteristic features of them are contained in the rate maps and can, for instance, be learned by machine-learning procedures. The knowledge gained in this way enables better and/or faster performance in recognition and localization tasks. It can, for example, be used by the system for attending to situation-specific spectral regions, for dynamic sensor adjustment, for suitable path planning in the context of scene exploration, or for active assessments of the quality of experience, such as in spaces for musical performance. This kind of feedback is knowledge-based and, consequently, reflective.

(i) Localizing Corrupted But Known Sounds

There is clear evidence for attentional effects in biological spatial hearing. It thus makes sense to apply top-down information in technological systems for sound localization as well. The experiment described in the following has addressed this issue by proposing a framework for binaural localization that exploits top-down knowledge about the spectral characteristics of different sources in acoustic scenes. A-priori information from source models is used to improve the localization process by selectively weighting interaural cues. The system, therefore, combines top-down and bottom-up information flows within a single computational framework. In detail, the following procedural steps have been taken in the experiments reported here (Two!Ears 2016, pp. 11–16); see also Ma and Brown (2015).

In the first step, rate maps are computed by the system for an acoustic mixture consisting of the target signal and interferer signals. Then those time-frequency features are estimated that are dominated by the target signal. The target signal was a 1.5-s speech sample, and the interferer signals were taken from the following set of sound samples: alarm, baby crying, drums, telephone ring, symphony. “Clean” power spectra of the target signal and each of the interferer signals were available from prior experiments. Then masks were built for

the enhancement of those components in the rate maps that are dominated by the target, and to penalize those that are dominated by the interferer. These steps were accomplished by machine learning.⁹ Figure 6 shows an example of the original rate map of the target/interferer mixture (upper panel) and the mask for weighting this rate map in a way that emphasizes the target signal (lower panel). Slight head rotations have therefore been applied to avoid front-back confusion (Ma et al. 2015; Ma and Brown 2016). The directional mapping was based on interaural level differences (ILDs) and interaural cross-correlation, both as a function of frequency.

“Sharpening of the ears” in the context discussed here denotes the fact that the auditory system has learned the relevant spectral pattern of predefined sound classes from the according to generalized rate-maps and, consequently, puts a higher weight on spectral components that fit these pattern while penalizing the remaining ones. This can be interpreted as a focusing of *attention* of the system on pattern elements that it has learned to be relevant in a particular situation—in other words, it is a case of reflective feedback. The experiment reported above thus shows that by exploiting learned rate-map-based source models sound-localization performance may improve substantially (Two!Ears 2016, pp. 11–15).

(ii) *Optimizing the Head Position in Multiple-source Scenarios*

As mentioned above, human beings tend to move their head into a direction where they either see the sound source or, alternatively, hear it best. This behavior also holds in situations with more than one sound source, for instance, in scenes with concurrent talkers. The listeners will then make a choice as to which talker to attend and concentrate on that one. This focusing results in a perceptual advantage for the “desired” talker—the so-called *Cocktail-party effect* (Cherry 1957). By putting the head in a favorable position for taking advantage of the cocktail-party effect, the listeners localize the desired talker and then segregate their speech signals from those of other talkers (Braasch et al. 2011; Parks and Braasch 2013). In the following, an algorithm is described that allows this reflective feedback to be included in auditory-system models. The algorithm requires three processing steps, namely, localization, head positioning, and segregation.

Experiments have been performed for various settings of two concurrent talkers in the horizontal plane. They revealed that best-positioning of the listener’s head with respect to the two talkers results in a perceptual advantage equivalent to a *signal-to-noise ratio* (SNR) of up to 30 dB (Deshpande and Braasch 2017). For finding the directions of the two talkers, the null-steering-antenna approach as described in Sect. 3.3 was utilized. Possible front/back confusions were solved by small head rotations. With the azimuth difference of the two talkers now known, the head was turned to a position in which the highest SNR was expected. This was done by looking-up in a table in which estimates of the SNR for all possible source positions had been stored beforehand as a-priori knowledge. The SNR values in the table were

⁹Conventional machine-learning methods usually work well for specific situations and tasks that they have been trained for, but they may unpredictably fail in situations that they are not prepared for. This risk is reduced with advanced machine-learning methods, such as DNN-based ones. These can refer to more abstract representations and are thus able to generalize to a certain extent.

collected with pink-noise. In the last step, the concurrent talker was suppressed, again by employing the null-steering-antenna method. All head rotations were simulated by virtual head rotation (Braasch et al. 2013). The overall results of the experiment support the hypothesis that turning the “better ear” toward the desired sound-source is a reasonable approach in most cases.

As to localization and/or suppression of a sound-source with the null-steering-antenna method, the following has to be noted. In environments with more than one source, such as with concurrent talkers, noise, and/or reverberation, there will be no complete null. Anyhow, if the signal components stemming from the identified source position are subtracted from the complete microphone signals, what is left stems from further sources, noise, or reverberation. Following this, the foreground and background can be separated, and the desired signal stands out more clearly. Consequently, in further processing steps, multiple sources can be identified, and dereverberation and denoising can be accomplished (Two!Ears 2016, pp. 33–39). If binaural signals are used as an input to the process, as was the case in the experiment reported above, it is advantageous to cancel the influence of the external ears, head, and torso by deconvolving the incoming signals with the adequate impulse responses¹⁰ before applying the cophase-and-subtract step.

Auditory Localization Aided by Sensorimotor Feedback

In the very beginning of this chapter, in Sect. 1, it was stated that the auditory system is an embedded component of the complete body and—due to communication of the body with the environment—closely bound to the environmental scene that the body is part of. Thus, any action of the body, such as auditory localization, can only be fully understood when inputs from all sensory modalities—including the coordinative ones with sensors for position, direction, and force—are taken into account. This, furthermore, also holds for the “world knowledge” that the body has incorporated. World knowledge requires intelligence within the body, yet, this intelligence is not solely represented in the “brain” of the system. For example, reflective motions of the system or of parts of it are triggered by the central nervous system, but the exact course of action follows (reflexive) patterns that are “known” at the stage where they are actually executed. Brooks (1991) has called this local phenomenon *intelligence without representation*. In O’Regan and Noe (2001) these ideas have been developed further in that these authors point to the tight integration of motion and sensory stimulation, a notion that complies with recent developments in embodied cognition. Namely, there it is postulated that our sensory experience arises from mastering sensorimotor contingencies, so to say, on the task of learning how stimuli vary as a function of bodily movement.

Along these lines of thinking a “sensorimotor-feedback” algorithm has been developed in Toulouse under the guidance of Patrick Danès. This algorithm exploits the correlation between streams of directional cues as delivered by the two ears of listeners and the accompanying motor actions that are attached to the sound-localization

¹⁰These impulse responses, also known as Head-Related Impulse Responses (HRIRs) vary with the direction of sound incidence and source distance. Their Fourier transforms are usually called Head-Related Transfer Functions (HRTFs).

processes (Bustamante and Danés 2017; Bustamante et al. 2017). The algorithm has been developed for use in robots. In principle, it can exploit world-knowledge, although the current implementation is restricted to the reflexive part of the feedback loop.

In the model architecture of the auditory system as depicted in Fig. 2, the sensorimotor level constitutes the lowest computational layer both in the model architecture at large and in the robot implementation. Since the respective processors in the robot must run under severe time and communication constraints, the current implementation does not yet entail any reflective ability.

Sensorimotor-feedback-aided sound-source localization is an active process that enables the disambiguation of front-back confusions, the determination of source distances, and other tasks by incorporating motor information. The sensorimotor localization strategy is organized into three layers (Bustamante et al. 2015)—as depicted in Fig. 7.

- **Stage A** performs instant estimates of the source azimuths and also detects the source activities from rate-map-like representations of small sliding time windows of the two ear-input signals. In this way, an initial probability-density function (pdf) of prospective source positions is formed
- **Stage B** assimilates this azimuth-over-time information and combines it with respective head-rotation motor commands in a stochastic filter. This filter generates a posterior probability-density function of the head-to-source directions
- **Stage C** provides a feedback controller that can move the head to improve the quality of localization based on suitable information-theoretic criteria. This leads to an improvement of the output of Stage B.

The stochastic filter in Stage B was set up as a state-space equation uniting the velocity-control vector of the head motion with the head-to-source directions, rendering a Gaussian-mixture approximation of the posterior pdf. The quality assessment in Stage C was performed via the entropy of the moment-matched approximation

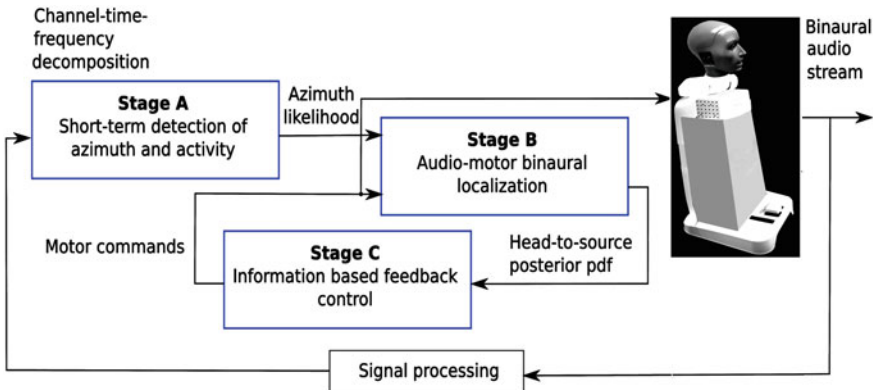


Fig. 7 Sensorimotor feedback (schematic)—adapted from Two!Ears (2015)

of the posterior pdf of the head-to-source directions. This three-stage sequence is repeated continuously while the localization process is going on. The algorithm has been formalized in statistical-signal-processing and information-theoretical terms. More details of it are given in Two!Ears (2015, pp. 56–67) and Two!Ears (2016, pp. 52–62).

In the implementation reported here, the auditory-modeling system was integrated into a robot which was designed to act as an agent that can explore its environment actively and independently, thereby taking on tasks like saving persons in Search-&-Rescue missions—see Blauert (2020), this volume, for an example performed in a virtual test environment.

Head Movements Controlled by Perceptual Congruence

There is ample evidence that auditory perception and understanding are subject to attention (Kaya and Elhilali 2016). Namely, auditory objects and scenes that human beings attend to are perceived and interpreted differently when they are in the focus of attention or not (Two!Ears 2014, pp. 43–55). These effects can be reflexive but also reflective with varying amounts of involved cognition. To understand the extent to which cognition may be involved, a functional hypothesis known as *Reverse-Hierarchy Theory* (RHT) (Hochstein and Ahissar 2002) provides useful hints. The idea of this hypothesis is as follows. As long as the relevance of a sound-source can be assumed to be due to primitive (and salient) perceptual features, attention will be paid to the source in a reflexive manner. However, if (and only if) cognition is required to analyze an auditory scene in order to determine which source may be relevant enough to attend to, reflective feedback will be employed. The RHT is thus also a statement of the economic behavior of organisms. As a practical example of this theory, the *Head-Turning Modulation* (HTM) model is introduced, which was conceptualized and developed by Cohen-L'hyver et al. (2015, 2016, 2020). The HTM algorithm comprises two major parts, namely,

1. A Dynamic Weighting model
2. A Multimodal-Fusion-&-Inference model.

The *Dynamic-Weighting Model* (DW) is able to control the movements of a head-and-torso simulator depending on whether a sound-source is regarded as *congruent* or not with respect to the current environment that the robot is about to explore. The notion of congruence is defined by the authors as a form of *Semantic Saliency* because it analyzes local singularities (Treisman and Gelade 1980) that represent perceived objects with respect to the environment in which they occur. However, as opposed to the traditional notion of saliency, which mainly relies on low-level cues of the signals perceived, the DW processes higher-level data using audio (and visual) classes the objects belong to. Dedicated classification experts provide this necessary information—see Cohen-L'hyver et al. (2020). Thus, the DW offers advanced reflective feedback regarding auditorily induced head rotation.

Congruence is computed on an environment-by-environment basis, relying on the fact that the same sound can be interpreted as “odd” in a particular space (such as a dog barking in a conference room) but completely “adequate” in another space (such

as a dog barking in a kennel). In the latter case, the barking would be predictable in a way. The DW algorithm now analyzes the pseudo-probability of inputs provided by classification experts to occur in the current environment. Thereby this algorithm, as well as the whole HTM model, does not rely on prior knowledge about the world being explored. Actually, the less predictable an object is, the more *important* it is rated. In this context, the amount of importance provides the reference for whether a head movement toward this object is justified or not. This decision, conferred to a robotic system together with the advantages listed above, enables the robot to direct its visual sensors towards the dedicated object properly and, consequently, improve the analysis of it.

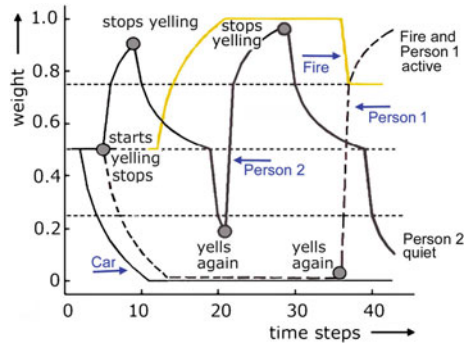
The *Multimodal-Fusion-&Inference Model* (MFI) also implements reflective feedback through the notion of autonomous enhancement of the knowledge perceived by the robot—the knowledge being the audiovisual classes brought by the classification experts. Mainly rooted in the principle of intrinsic motivation (Berlyne 1950), the MFI is mainly constituted by a self-supervised online learning algorithm, in charge with learning the link between the audio and visual information brought by the classifiers during the exploration, coupled to the ability to assess by itself how much the knowledge it gained from learning is of sufficient quality to be trusted by the DW. Indeed, the DW relies on a robust audiovisual representation of the world in order to compute the congruence of the objects detected in it in a meaningful and relevant manner. As for the DW, the MFI extensively uses head movements to feed its learning algorithm with complete audiovisual data. Also, as for the DW, the MFI aims at inhibiting irrelevant or unnecessary head movements.

Both modules, DW and MFI, provide reflective feedback loops between the information perceived by the different sensors of the robot and their respective elements for analysis, namely, classification experts and the triggering of head movements. These loops are highly adaptive since they are adjusted by the HTM system itself all along with the life of the robot.

Figure 8 illustrates an example where the notion of congruence versus incongruence has been applied. The assigned weights are measures of the assumed urgency for turning the head into the direction of the respective sound sources. The weighting process in the depicted 4-source scenario with two persons, a car, and a fire, proceeds as follows. The car is started and then parked close to two persons. Its motor stays running. At Timestep 5, both persons begin yelling, but while Person 1 stops immediately, Person 2 keeps on until about Step 10 and only then stops. At Step 20, Person 2 starts to yell again and does so until about Step 25 to stay quiet from then on. At about Timestep 10, a fire breaks out and produces loud crackling noise. At Step 35, Person 1 begins to yell again and stays on doing so.

The following effects can be observed from the trajectories of the weights. The weight assigned to a source decreases when it falls silent or continues to emit the same sound for a longer period. The decrease is not abrupt but follows a shallow slope. The weight increases, however, when a source starts to send from scratch or starts sending modified or new sounds. Incongruence has obviously something to do with the information that a sound source provides. So far in the system reported here, a cross-correlation criterion rated the amount of incongruence.

Fig. 8 Example of congruence-driven weight assignment in a dynamic 4-source scenario—adapted from Cohen-L’hyver (2017)



The semantic content of the emitted sound has not yet been considered. For example, information of the following kind is not exploited in the current version of the system. Is a sound human speech and, if yes, is it female or male speech? Does it originate from a known or unknown person and, what does this person actually express in terms of semantic content? Recent progress in speech-dialog systems—see Sect. 4.2—will support further development. Further ideas along these lines, including the use of visual cues, are discussed in the chapter by Cohen-L’hyver et al. (2020), this volume.

4 Cognitive Feedback

4.1 Feedback Processes in the “Brain” of the Auditory System

The need for cognitive feedback is motivated by the following statement.

Human beings do not react according to what they perceive, but rather, they react on the grounds of what the percepts mean to them in their current action-specific, emotional and cognitive situation—a much more complex process than just perception

A model of the auditory function that also includes cognitive processes and reflective interaction with the world needs a component that incorporates world knowledge and cognitive reasoning, among other capabilities. This component must, for instance, be able to form objects from proto-events, analyze auditory scenes—thereby taking cross-modal cues into account (e.g., visual and/or tactile ones)—and assign meaning to objects. In other words, a component that represents, so to say, the “brain” of the model system. *Blackboard systems* (Engelmore and Morgan 1988) are one possibility to realize the functions laid out above. In the model architecture depicted in Fig. 2, this component is composed from the actual *blackboard* together with the *experts* and the *scheduler*. Architecture and functions of a specific system of this kind are described in more detail in Schymura and Kolossa (2020), this volume. In that system,

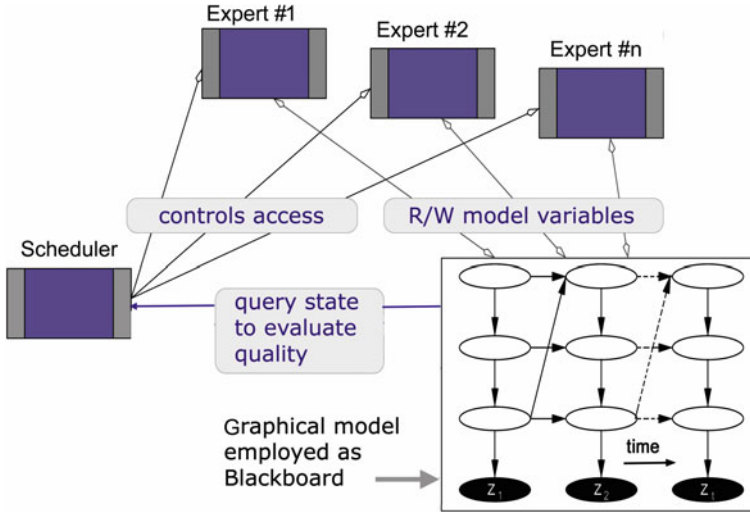


Fig. 9 Interaction of blackboard and experts under control of a Scheduler—adapted from Blauert et al. (2014)

feedback from higher levels is integrated by using graphical models (Murphy 1998) as the active blackboard architecture. Higher-level processes in application-specific subsystems, such as a software expert on scene analysis, can set variables according to their particular intentions. Then, after an inference in the graphical model has been carried out accordingly, it becomes clear how higher-level feedback corresponds with the rules and observations of the system, and what implications can eventually be drawn from it (Fig. 9).

The graphical models stem from multiple sources of information and are composed on the blackboard to form one comprehensive description of a specific auditory, audio-visual, or audio-visual-tactile scene. The model parameters are adjusted in order to create a world model, namely, a description of the state of the environment which optimally matches all observations, that is, all sensor data that have been made available to the cognitive system.

This structure allows for many types of feedback, which can be initiated whenever the output of the system is not sufficiently reliable. Insufficient reliability is detectable within the graphical models, but care must be taken to distinguish continuous-valued variables like locations or intensities from discrete-valued variables like spoken words or source identities. If there is insufficient reliability in a continuous-valued variable, this can be seen from large variances of the estimate. For instance, if the system is tracking an acoustic source, high variances of the location estimate are indicative of an unreliable interpretation, as has been successfully exploited in Schymura et al. (2014). For discrete-valued variables, confusions are detected when multiple interpretations are assigned high likelihoods. One example where such problems can occur is given by situations with conflicting evidence, that is, when different

subsystems assign two or more contradictory interpretations high likelihoods. For example, one source may be interpreted as a human talker by the acoustic model but as a speech-emitting radio by the visual model.

In both types of confusion, continuous-valued and discrete-valued, the graphical model architecture of the blackboard helps trigger feedback and disambiguation. More specifically, the blackboard system takes advantage of the connectivity of the complete graphical model for this purpose. That is, when a variable on the blackboard is shown to have a high degree of uncertainty, the underlying causes of uncertainty are traced by following the dependency relationships of the variable backwards. The results of these processes are then communicated in a top-down manner to other stages of the auditory-system model, for instance, to initiate suitable adaptation processes, or to trigger motor actions, such as goal-oriented hand and body motions.

4.2 Understanding Auditory Agents

In order to be able to control and assess advanced auditory-system models, such as the one sketched in Fig. 2, some kind of communication channel between the model and the outside world has to be established. This way users can understand what is happening inside the systems. For example, (a), what is the current world model that a system is using and/or, (b), what are the action plans that it is pursuing and actually carrying out?

- (a) Answers to the first query require a way for reporting from inside the “brain” of such a system. This communication can be achieved by applying advanced text- or speech-dialog systems—see, for example, Jurafsky and Martin (2017), and Möller (2010). Recently, these systems have enjoyed substantial progress, and they are now successful items of consumer technology, among other usages.
- (b) For answering the second query, an obvious method is to observe the actual actions of the system while in use and then to analyze and evaluate these actions—for instance, with respect to intended purposes.

For observation and assessment of the behavior of robots that are controlled by a model system as depicted in Fig. 2, a dedicated (virtual) test environment has been developed—for details see Blauert (2020), this volume. This test environment comprises, among further ones, the following functions¹¹:

- *Environment*. This class provides the basis of the virtual test environment. Controlled by computer commands, it is able to generate auditory scenarios of moderate complexity. For example, the current version generates rectangular spaces in which sound sources are placed, and in which a robot can move about freely
- *Sound-sources*. Here basic information is deposited regarding each sound-source used in the scenario, such as source positions, identities (names) and test sounds

¹¹These classes were originally programmed in MATLAB®.

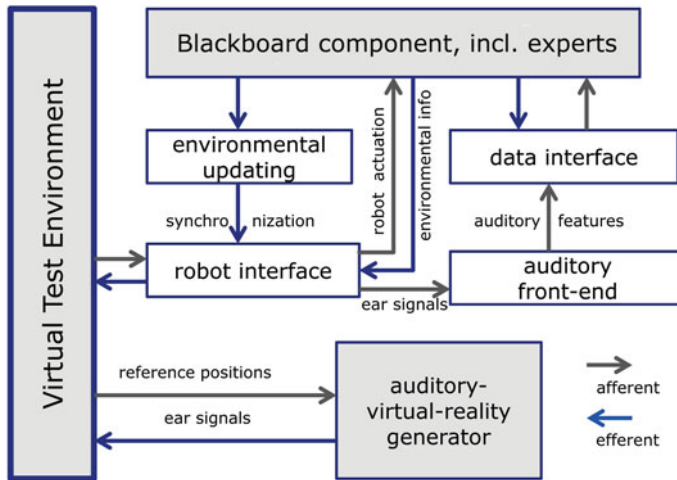


Fig. 10 Schematic for the test environment for advanced binaural-system models

to be emitted. The positions of the sound sources, as well as their emission type and timing, are communicated from here to the environment class

- *Robot Control.* In this class, the virtual representation of the robot is stored and administered. It admits access of the test-environment system to the robot platform and is in control of movements of this platform
- *Artificial Head.* On top of the robotic platform, an artificial head is mounted, which is movable with respect to the platform. Control of head position and movement is carried out via this class. Further, the head-related transfer functions (HRTFs) as needed for auralization, are applied here.

The test environment connects seamlessly with its blackboard component and the experts attached to it—see Fig. 10 for a schematic plot of the arrangement. This enables the experts to commit motor orders to the robotic platform with its movable head-and-torso simulator on top. Environmental information, as provided by the experts, will be fed back to the test environment accordingly, also through this interface. A dedicated synchronization expert provides synchrony between the test environment and the blackboard. Because the auditory virtual-reality generator is directly controlled by the test environment, this unit, among other sound signals, also generates the ear signals for the (virtual) robots. These are then sent to the peripheral components of the auditory-system model, namely, its “auditory front end”. Here preprocessing of the ear signals is performed, and relevant auditory features are extracted. These are consequently forwarded to the blackboard component and thus also to the experts. The virtual test environment allows for quick and reliable monitoring of basic feedback mechanisms in the auditory-system model.

5 Concluding Remarks

While research efforts into the human auditory system have traditionally concentrated on bottom-up processes, the focus has now broadened, and top-down processes have caught the attention of researchers as well. Since bottom-up and top-down processes are intimately linked, it is evident that they will form feedback loops within the auditory system, and with regard to the communication of the auditory system with other parts of the body, and with the environment. This fact makes auditory feedback a relevant target also for applied research and engineering approaches, for instance, in human/machine communication, and robotics. In this chapter, some basic problems and research efforts regarding this area have been introduced and discussed. The majority of the examples have been taken from the recent EU-Project TWO!EARS.¹²

Acknowledgements The work reported here has been supported by the EU-FET project TWO!EARS (Contract No. 618075). This chapter contains excerpts from the respective project reports. The authors are indebted to those of our colleagues in this project who contributed to the work on auditory feedback, in particular, S. Argentieri, G. Bustamante, B. Cohen-L'hyver, P. Danès, Th. Forgue, B. Gas, Y. Guo, Y. Kashef, R. Kim, D. Kolossa, A. Kohlrausch, N. Ma, J. Mohr, A. Podlubne, Chr. Schymura, Th. Walther, and H. Wierstorf. Further, they thank two anonymous reviewers for valuable comments and suggestions.

References

- Berlyne, D.E. 1950. Novelty and curiosity as determinants of exploratory behavior. *British Journal of Psychology* 41 (1–2): 68–80.
- Blauert, J. 1997. *Spatial Hearing—The Psychophysics of Human Sound Localization*, 2nd revised ed. Cambridge, MA and London, UK: The MIT Press.
- Blauert, J. 2020. A virtual testbed for binaural agents. In *The Technology of Binaural Understanding*, eds. by Blauert, J. and Braasch, J. Springer and ASA Press.
- Blauert, J., J. Braasch, J. Buchholz, H. Colburn, U. Jekosch, A. Kohlrausch, J. Mourjopoulos, V. Pulkki, and A. Raake. 2010. Aural assessment by means of binaural algorithms—the AabbA project. In *Binaural processing and spatial hearing, proceeding of 2nd international symposium on auditory and audiological research – ISAAR'09*, eds. Buchholz, J., T. Dau, J. Dalsgaard, and T. Poulsen, 113–124. DK-Ballerup: The Danavox Jubilee Foundation.
- Blauert, J., D. Kolossa, and P. Danès. 2014. Feedback loops in engineering models of binaural listening. In *Proceedings of Meetings on Acoustics (POMA)*, vol. 21, 1–11.
- Blauert, J., K. Obermayer. 2012. Rückkopplungswege in Binauralmodellen—(Feedback loops in binaural models). In *Fortschr. Akustik, DAGA, 2012. 2015–2016*. D-Oldenburg, Deutsche Gesellschaft für Akustik.
- Blauert, J (ed). 2017. Reading the world with two ears. In *Proceedings 24th International Congress on Sound and Vibration (ICSV)*, London, Intl. Inst. of Acoustics & Vibration (IIAV), Auburn AL, USA. www.iiav.org/archives_icsv_last/2017_icsv24/index.html. Accessed 20 Aug 2019.
- Borg, E., and S.A. Counter. 1989. The middle ear muscles. *Scientific American* 261 (2): 74–78. <https://doi.org/10.1038/scientificamerican0889-74>.

¹²Further details of the results of this project, including its open-software repository, can be accessed through the project's website <http://www.twoears.eu> [last accessed August 18, 2019] and from GitHub under <https://github.com/TWOEARS> [last accessed August 20, 2019].

- Braasch, J., S. Clapp, A. Parks, T. Pastore, and N. Xiang. 2013. A binaural model that analyses aural spaces and and stereophonic reproduction systems by utilizing head movements. In *The Technology of Binaural Listening*, ed. Blauert, J., 201–223. Springer and ASA Press. https://doi.org/10.1007/978-3-642-37762-4_1.
- Braasch, J., A. Parks, and N. Xiang. 2011. Utilizing head movements in the binaural assessment of room acoustics and analysis of complex sound source scenarios. *Journal of the Acoustical Society of America* 129: 2486.
- Bregman, A.S. 1990. *Auditory Scene Analysis—The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.
- Brooks, R. 1991. Intelligence without representation. *Artificial Intelligence* 47: 139–159.
- Brown, G.J., and M. Cooke. 1994. Computational auditory scene analysis. *Computer Speech and Language* 8 (4): 297–336.
- Brown, G.J., R.T. Ferry, and R. Meddis. 2010. A computer model of auditory efferent suppression: Implications for the recognition of speech in noise. *Journal of the Acoustical Society of America* 127 (2): 943–954.
- Brown, M., R. Venecia, and J. Guinan Jr. 2003. Responses of medial olivocochlear neurons. *Experimental Brain Research* 153 (4): 491–498. <https://doi.org/10.1007/s00221-003-1679-y>.
- Bustamante, G., and P. Danès. 2017. Multi-step-ahead information-based feedback control for active binaural localization. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6729–6734. <https://doi.org/10.1109/IROS.2017.8206589>.
- Bustamante, G., P. Danès, T. Forgue, and A. Podlubne. 2016. Towards information-based feedback control for binaural active localization. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6325–6329.
- Bustamante, G., P. Danès, T. Forgue, A. Podlubne, and J. Manhès. 2017. An information based feedback control for audio-motor binaural localization. In *Autonomous Robots 2017, Special Issue on Active Perception*, ed. G. Sukhatme, 1–14. Berlin: Springer. <https://doi.org/10.1007/s10514-017-9639-8>.
- Bustamante, G., A. Portello, and P. Danès. 2015. A three-stage framework to active source localization from a binaural head. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'2015)*, Brisbane, Australia.
- Celesia, G., and G. Hickok. eds. 2015. *Handbook of Clinical Neurology. Vol. 129: The Human Auditory System*. Edinburgh, New York: Elsevier.
- Cherry, E.C. 1957. *On Human Communication*. Cambridge, MA: The MIT Press.
- Cohen-L'hyver, B. 2017. Modulation de mouvements de tête pour l'analyse multimodale d'un environnement inconnu (modulation of the head movements by multimodal analysis of an unknown environment). Ph.D. thesis, Paris: University Pierre and Marie Curie.
- Cohen-L'hyver, B., S. Argentieri, and B. Gas. 2015. Modulating the auditory turn-to-reflex on the basis of multimodal feedback loops: the dynamic weighting model. in *IEEE Robio 2016 - International Conference on Robotics and Biomimetics*, 1109–1114.
- Cohen-L'hyver, B., S. Argentieri, and B. Gas. 2016. Multimodal fusion and inference using binaural audition and vision. In *International Congress on Acoustics (ICA 2016)*.
- Cohen-L'hyver, B., S. Argentieri, and B. Gas. 2020. Audition as a trigger of head movements. In *The Technology of Binaural Understanding*, eds. J. Blauert and J. Braasch, 697–731. Cham, Switzerland: Springer and ASA Press.
- Corbetta, M., G. Patel, and G. Shulman. 2008. Review: The reorienting system of the human brain: From environment to theory of mind, 306–324. <https://doi.org/10.1016/j.neuron.2008.04.017>.
- Corkill, D.D. 1991. Blackboard systems. *Artificial Intelligence Review* 2 (2): 103–118. <https://doi.org/10.1007/BF00140399>.
- de Andrade, K.C., E.D. Camboim, A. Soares Ido, M.V. Peixoto, S.C. Neto, and P. de Lemos Menezes. 2011a. The importance of acoustic reflex for communication. *American Journal of Otolaryngology* 32 (3): 221–227.
- de Andrade, K.C.L., S.C. Neto, and P. de Lemos Menezes. 2011b. The importance of acoustic reflex in speech discrimination. In *Speech Technology*, ed. I. Ipsic (Chap. 9: 185–194). London: Intech.

- Deshpande, N., and J. Braasch. 2017. Blind localization and segregation of two sources including binaural head movements. *Journal of the Acoustical Society of America* 142: EL113–EL117. <https://doi.org/10.1121/1.4986800>.
- Djupesland, G. 1964. Middle-ear muscle reflexes elicited by acoustic and nonacoustic stimulation. *Acta Oto-Laryngologica Suppl.* 188: 287–292.
- Durlach, N.I. 1960. Note on the equalization and cancellation theory of binaural masking level differences. *Journal of the Acoustical Society of America* 32: 1075–1076.
- Engelmore, R., and S. Morgan (eds.). 1988. *Blackboard Systems*. Boston, MA: Addison-Wesley.
- Erman, L.D., F. Hayes-Roth, V.R. Lesser, and D.R. Reddy. 1980. The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys* 12 (2): 213–253. <https://doi.org/10.1145/356810.356816>.
- Fay, R., and A. Popper (eds.). 1992–2017. *Springer Handbook of Auditory Research*. Berlin: Springer.
- Ferry, R.T., and R. Meddis. 2007. A computer model of medial efferent suppression in the mammalian auditory system. *Journal of the Acoustical Society of America* 122 (6): 3519–3526.
- Ghitza, O., D. Messing, L. Delhorne, L. Braida, E. Bruckert, and M. Sondhi. 2007. Towards predicting consonant confusions of degraded speech. In *Hearing – from Sensory Processing to Perception*, eds. Kollmeier, B., G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey, 541–550. New York, NY: Springer.
- Goodfellow, I., Y. Bengio, and A. Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: The MIT Press. <http://www.deeplearningbook.org>. Accessed 12 July 2019.
- Guinan, J.J. 1996. Physiology of olivocochlear efferents. In *The Cochlea*, eds. Dallos, P., A.N. Popper, and R.R. Fay, 435–502. Berlin: Springer.
- Guinan jr, J., 2010. Cochlear efferent innervation and function. *Curr. Opin. Otolaryngol. Head Neck Surg.* 18 (5): 447–453. <https://doi.org/10.1097/MOO.0b013e32833e05d6>.
- Harkrider, A.W., and C.D. Bowers. 2009. Evidence for a cortically mediated release from inhibition in the human cochlea. *Journal of the American Academy of Audiology* 20 (3): 208–215.
- He, J., and Y. Yu. 2009. Role of the descending control in the auditory pathway. In *Oxford Handbook of Auditory Science, Vol. 2: The Auditory Brain*, eds. Rees, A., and A.R. Palmer, 247–268. New York, NY: Oxford University Press.
- Hochstein, S., and M. Ahissar. 2002. View from the top: Hierarchies and reverse hierarchies review. *Neuron* 36 (3): 791–804.
- Holdsworth, J., R. Patterson, I. Nimmo-Smith, and P. Rice. 1988. *Implementing a gammatone filter bank*. In: *Annex C of the SVOS Final Report*. Cambridge, UK: University of Cambridge.
- ISO. 2003. Iso 226:2003: Acoustics—normal equal–loudness-level contours (International Organisation for Standards). <https://www.iso.org/standard/34222.html>. Accessed 4 Aug 2019.
- Jekosch, U. 2005. Assigning of meaning to sounds—Semiotics in the context of product-sound design. In *Communication Acoustics*, ed. J. Blauert, 193–221. Berlin: Springer.
- Jurafsky, D., and J.K. Martin. 2017. Advanced dialog systems. In *Speech and Language Processing* (Chap. 30). London, UK: Pearson. <https://web.stanford.edu/~jurafsky/slp3/30.pdf>.
- Jurgens, T., N.R. Clark, W. Lecluyse, and R. Meddis. 2016. Exploration of a physiologically-inspired hearing-aid algorithm using a computer model mimicking impaired hearing. *International Journal of Audiology* 55 (6): 346–357.
- Kaya, E.M., and M. Elhilali. 2016. Modelling auditory attention. *Philosophical Transactions of the Royal Society B* 327 (372): 20160099. <https://doi.org/10.1098/rstb.2016.0101>.
- Kung, B.C., and T.O. Willcox Jr. 2007. Examination of hearing and balance. In *Neurology and Clinical Neuroscience*, eds. A.H.V. Schapira and E. Byrne (Chap. 25: 318–327). Philadelphia: Mosby Elsevier. <https://doi.org/10.1016/B978-0-323-03354-1.50029-8>.
- Leng, S., W. Ser, C. Ko. 2008. A simple constrained based adaptive null steering algorithm. In *Proceeding of EUSIPCO 2008, EURASIP, Lausanne*.
- Longtin, A., and J.R. Derome. 1989. A new model of the acoustic reflex. *Biological Cybernetics* 53 (5). <https://doi.org/10.1007/BF00336564>.

- Lopez-Poveda, E.A., A. Estaquio-Matin, J.S. Stohl, R.D. Wolford, R. Schatzer, J.M. Gorospe, S. Ruiz, F. Benito, and B.S. Wilson. 2017. Intelligibility in speech maskers reflex binaural cochlea implantant sound coding strategy inspired by the contralateral medial olivocochlear reflex. *Hearing Research* 348: 134–137. <https://doi.org/10.1016/j.heares.2017.02.003>.
- Lopez-Poveda, E.A., A. Estaquio-Matin, J.S. Stohl, R.D. Wolford, R. Schatzer, and B.S. Wilson. 2016. A binaural cochlear implant coding strategy inspired by the contralateral medial olivocochlear reflex. *Ear & Hearing* 37 (3): e138–e148.
- Lutman, M.E., and A.M. Martin. 1979. Development of an electroacoustic analogue model of the middle ear and the acoustic reflex. *Journal of Sound and Vibration* 64 (1): 133–157.
- Ma, N., and G.J. Brown. 2015. Robust localization of multiple speakers exploiting deep neural networks and head movements. In *Proceeding of Interspeech*, 3302–3306.
- Ma, N., and G.J. Brown. 2016. Speech localisation in a multitalker mixture by humans and machines. In *Proceeding of Interspeech 2016* (International Speech-Communication Association (ISCA)), 3359–3363. <https://doi.org/10.21437/Interspeech.2016-1149>.
- Ma, N., T. May, and G.J. Brown. 2015. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (12): 2444–2453.
- Mercedes. 2002. Pre-safe sound. <https://www.mercedes-benz.com/en/mercedes-benz/next/connectivity/pre-safe-sound-playing-pink-noise-in-the-split-second-before-impact/>. Accessed 9 Aug 2019.
- Mi, J., M. Groll, and H.S. Colburn. 2017. Comparison of a target-equalization-cancellation approach and a localization approach to source separation. *Journal of the Acoustical Society of America* 142: 2933–2941. <https://doi.org/10.1121/1.5009763>.
- Møller, A. 1974. The acoustic middle ear muscle reflex. In *Auditory System*, ed. H. Ades et al., 520–566. Berlin, Heidelberg: Springer.
- Møller, A.R. 1962. Acoustic reflex in man. *Journal of the Acoustical Society of America* 35: 1524. <https://doi.org/10.1121/1.1918384>.
- Möller, S. 2010. *Quality of Telephone-based Spoken Dialogue Systems*. Berlin: Springer. <https://doi.org/10.1007/b100796>.
- Moore, B. 1989. *An Introduction to the Psychology of Hearing*, 3rd ed. New York: Academic.
- Moore, B.C.J. (ed.). 1995. *Handbook of Perception and Cognition. Vol.6: Hearing*. New York: Academic.
- Mukerji, S., A.M. Windsor, and J.D. Lee. 2010. Auditory brainstem circuits that mediate the middle ear muscle. *Trends in Amplification* 14 (3): 170–191. <https://doi.org/10.1177/1084713810381771>.
- Murphy, K. 1998. A brief introduction to graphical models and bayesian networks. In *USB Computer Science Tutorials*. University of California, Berkeley. <https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html>. Accessed 8 Aug 2019.
- O'Regan, J., and A. Noe. 2001. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* 25 (5): 939–1031.
- Parks, A., and J. Braasch. 2013. Head tracking and source localization in reverberant cocktail-party scenarios. *Journal of the Acoustical Society of America* 133 (5): 3511.
- Perrett, S., and W. Noble. 1997. The contribution of head motion cues to localization of low-pass noise. *Perception and Psychophysics* 59: 1018–1026.
- Petersen, S., and M. Posner. 2012. The attention system of the human brain: 20 years after. *Annual Review of Neuroscience* 21 (35): 73–89. <https://doi.org/10.1146/annurev-neuro-062111-150525>.
- Plack, C.J. (ed.). 2010. *Oxford Handbook of Auditory Science: Hearing*. Oxford, UK: Oxford University Press.
- Raake, A., and J. Blauert. 2013. Comprehensive modeling of the formation process of sound-quality. In *Proceeding of 5th International Workshop on Quality of Multimedia Experience (QoMEX, Klagenfurt)*, 76–81.

- Schofield, B.R. 2009. Structural organization of the descending auditory pathway. In *Oxford Handbook of Auditory Science, Vol. 2: The Auditory Brain*, eds. A. Rees, and A. R. Palmer, 43–64. New York, NY: Oxford University Press.
- Schymura, C., and D. Kolossa. 2020. Blackboard systems for cognitive audition. In *The Technology of Binaural Understanding*, eds. Blauert, J., and J. Braasch, 91–111. Cham, Switzerland: Springer and ASA Press.
- Schymura, C., T. Walther, D. Kolossa, N. Ma, and G. Brown. 2014. Binaural sound source localisation using a bayesian-network-based blackboard system and hypothesis-driven feedback. In *Proceeding of Forum Acusticum 2014*, European Acoust. Ass. (EAA), Kraków, Poland.
- Shamma, S. 2013. Cortical processes for navigating complex acoustical environments (invited talk). In *3rd Binhaar Symp.*, CNRS/LAAS- Univ. de Toulouse Paul Sabatier.
- Smith, D.W., R.K. Aouad, and A. Keil. 2012. Cognitive task demands modulate the sensitivity of the human cochlea. *Frontiers in Psychology* 3, Article 30. <https://doi.org/10.3389/fpsyg.2012.00030>.
- Sutojo, S., S. Van de Par, J. Thiemann, and A. Kohlrausch. 2020. Auditory Gestalt rules and their application. In *The Technology of Binaural Understanding*, eds. Blauert, J. and J. Braasch, 33–59. Cham, Switzerland: Springer and ASA Press.
- Treisman, A.M., and G. Gelade. 1980. A feature-integration theory of attention. *Cognitive Psychology* 12 (1): 97–136. [https://doi.org/10.1016/0010-0285\(80\)90005-5](https://doi.org/10.1016/0010-0285(80)90005-5).
- Two!Ears. 2014. Feedback-loop selection and listing, part B: literatur survey. In *Two!Ears Publications*, ed. Walther, T. (Chap. Project deliverables, item). <https://doi.org/10.5281/zenodo.2595244>.
- Two!Ears. 2015. Specification of feedback loops and implementation progress. In *Two!Ears Publications*, ed. J. Blauert and T. Walther, 56–61 (Chap. Project deliverables, item d4.2). <https://doi.org/10.5281/zenodo.2595224>.
- Two!Ears. 2016. Final integration-&-evaluation. In *Two!Ears Publications*, eds. J. Blauert and T. Walther (Chap. Project deliverables, item d4.3). <https://doi.org/10.5281/zenodo.2591202>.
- Yost, W. 2007. *Fundamentals of Hearing: An Introduction*, 5th ed. San Diego, New York, London: Elsevier.
- Zhao, W., and S. Dhar. 2011. Fast and slow effects of medial olivocochlear efferent activity in humans. *PLoS One* 6 (4): e18725. <https://doi.org/10.1371/journal.pe.0018725>.

Auditory Gestalt Rules and Their Application



Sarinah Sutojo, Joachim Thiemann, Armin Kohlrausch
and Steven van de Par

Abstract The formation of *auditory objects* is of high interest for both the understanding of human hearing as well as for computer-based analysis of sound signals. Breaking down an acoustic scene into meaningful units facilitates the segregation and recognition of sound-sources from a mixture. These are abilities that are particularly challenging for machine listening as well as for hearing-impaired listeners. An early approach to explaining object perception in the visual domain was made by the Gestalt psychologists. They aimed at setting up specific rules according to which sensory input is grouped into either one coherent or multiple separate objects. Inspired by these Gestalt Rules and by exploiting physical and perceptual properties of sounds, different algorithms have been designed to segregate sound mixtures into auditory objects. This chapter reviews some literature on such algorithms and the underlying principles of auditory object formation with a special focus on the connection between perceptual findings and their technical implementation.

1 Introduction

A single speaker, an instrument playing, or a passing car, each create a combination of different sound components that a listener may easily identify as emerging from one specific object. Despite the different nature of the sound components and their dispersion over time and frequency, the separate components share enough evidence for the human auditory system to assign them to a common source of origin. In computer-based analysis of audio contents, organizing the sound components that occur in an auditory scene into meaningful elements, remains one of the main challenges.

S. Sutojo · J. Thiemann · S. van de Par (✉)
Department of Medical Physics and Acoustics, Cluster of Excellence Hearing4All,
University of Oldenburg, 26111 Oldenburg, Germany
e-mail: steven.van.de.par@uni-oldenburg.de

A. Kohlrausch
Human-Technology Interaction, Eindhoven University of Technology,
5600 MB Eindhoven, The Netherlands

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_2

An *auditory object* describes such a meaningful element. It either relates to a physical object or to the percept that goes with it. By means of object *formation* and *selection*, the multitude of incoming sounds can be reduced to components conveying relevant information about the environment, and eventually direct the listener's attention to these meaningful events. Not only does this make the analysis of inputs more purposeful and efficient, but considering limited processing capacities, the restriction to relevant elements is also a necessary step to enable handling of complex situations.

When gaining information from an auditory scene, both machines and human listeners have access to physical cues that are inherent to the acoustic signal such as frequency content or onset time across frequency bands. The challenge lies in the extraction and interpretation of these attributes in order to determine whether specific sound components should be grouped or separated. While the human auditory system exhibits a high degree of flexibility in unknown environments, the computer algorithms still lack in terms of proper selection and interpretation of physical cues in order to achieve the same extent of robustness in situations with an increasing number and variety of sources or in reverberant environments. Despite the broadened technical possibilities and the insight on human perception that has been gained in recent years, a comprehensive understanding of the information processing as it takes place in the auditory system is not yet available. However, there are multiple applications for computational methods that allow to reliably segregate sound sources. From automatic speech-recognition systems, which often require a front-end that can robustly distinguish between target speech and background (Narayanan and Wang 2013a; Wang and Wang 2016), to hearing aids, and music analysis, the automatic formation of meaningful objects is of substantial interest.

In this chapter, the concept of the *auditory objects* and the principles that underlie their formation, are regarded from both the perceptual perspective, as well as from an application-oriented view. The field of study that is concerned with the understanding of the perceptual processes in complex auditory situations is termed *auditory-scene analysis* (ASA) (Bregman 1990). Summarized under the term *computational auditory-scene analysis* (CASA) (Wang and Brown 2006) is the research field guided towards the technical implementation of scientific findings on ASA.

Figure 1 outlines the two views on *auditory objects* and some of the essential aspects that are addressed in this chapter. In the terminology of ASA, summarized on the left-hand side of the diagram, an auditory object is referred to as a *stream* (Bregman 1990) that defines the perception of different sound components as a coherent whole. From the CASA perspective, represented on the right-hand side of Fig. 1, the auditory object can be seen as a cluster of spectrotemporal units that are dominated by the same acoustic source and can thus be used to reconstruct the perceptual (auditory) components evoked by this source. A number of object-formation principles and constraints in the auditory system were derived from empirical studies applying stimuli that allowed for manipulation of physical signal properties and permitted the observation of their influence on stream perception. The Gestalt Rules are a set of principles stemming originally from the visual domain and allowing the definition of objects (cf. Binder et al. 2009). These principles have then also inspired

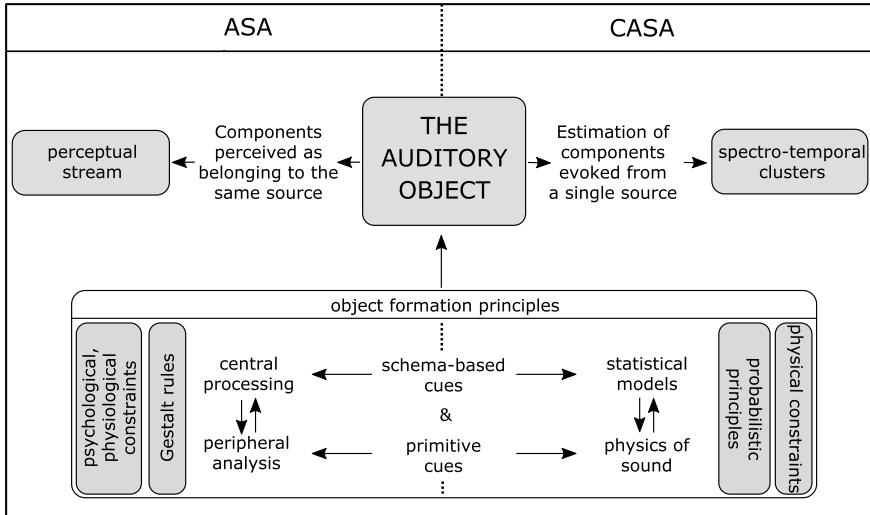


Fig. 1 The *Auditory Object* and its formation principles and grouping cues, regarded from both the perceptual (ASA) and application-oriented (CASA) perspective

ASA concepts. Although the original formulation of the Gestalt principles was rather philosophically oriented, some of them have evolved into more specific ones in terms of physical features (Jäkel et al. 2016) and thus became technically realizable with recent signal-processing methods.

The rules behind auditory-stream perception can be divided into two stages—see also the bottom half of the diagram. On the one hand, the inherent physical properties of the input signal serve as a basis for object formation and are summarized under *primitive grouping cues* (Bregman 1990). Regarding the auditory system, most of these are already represented in the peripheral analysis and are considered to influence the perception in a type of bottom-up process. On the other hand, the prior knowledge and a more conscious decision by the listeners determine which signal components the listeners attend to and how these components are assembled and recognized, for example, as a familiar acoustic source. These processes are referred to as *schema-based cues* and are considered to affect the perception in a top-down way. Rather than being purely determined by the stimulus input, these cues are driven by higher cognitive processes such as attention and prior knowledge. An essential part of ASA is the design of physiologically plausible models that reconstruct known stages in human auditory processing and attempt to precisely reproduce data from human-performance studies (Beauvois and Meddis 1991, 1996; McCabe and Denham 1997).

While the ASA field investigates which features humans use for stream segregation and the perceptual consequences that go with it, the CASA research field

(displayed on the right of the diagram), is more concerned with the question of which physical properties are extractable from the acoustic input signals, and how these can be put to use in algorithms for extracting information from an auditory scene, or even reconstructing individual sources. The equivalent to the perceptual stream in CASA is the clustering of signal fragments that are likely to be evoked by the same source. Several findings and terminologies that are used in ASA are translated into physically and statistically manageable quantities to make them suitable for technical implementations (Brown and Cooke 1994; Hu and Wang 2006). Consequently, the grouping cues are either obtained from an analysis of the acoustic input signal (primitive cues), or they have to be provided by pre-trained statistical models (schema-based cues). CASA algorithms do not necessarily need to be physiologically plausible or restricted to the limits of the auditory system but are often inspired by human processing strategies. This chapter deals with systems that provide inputs from only two microphones, similar to the two ears that humans use for ASA.

The intention of this chapter is to give an overview of recent findings on object formation in human hearing and provides some examples of computational methods that imitate these human abilities. In Sect. 2, the concept of an auditory stream is described along with the basic ideas on decomposing auditory scenes with binary masks such as used in CASA. In Sect. 2.1, feature types and grouping principles are reviewed by looking separately at primitive and schema-based cues. In the conclusion, an algorithm topology is suggested for the integration of various feature types with the help of similarity estimations.

2 The Auditory Object

An object in the environment is experienced as a defined entity, exhibiting characteristic features and standing in distinct relationships with other objects. For the visual domain, so-called *Gestalt Rules* have been formulated, which provide a more specific definition of what is perceived as a coherent object and what attempts are useful to capture regularities between stimuli and the resulting percepts in an auditory scene (cf. Jäkel et al. 2016). In the context of this chapter, these Gestalt Rules are looked at with regard to auditory objects, for which they also apply in a slightly modified form. From the viewpoint of auditory signal processing (e.g., in CASA), the process of object formation is approached by the search for components which are evoked by the same acoustic source. It is usually aimed at estimating an ideal binary mask (IBM) (Wang and Brown 2006) which labels spectrotemporal units that are dominated by the considered source.

2.1 Auditory Streams

Objects in auditory perception are referred to as *streams*. A stream defines a group of successive or simultaneous sound components as a coherent whole that appears to originate from the same source. *Streaming* describes the processes which influence how many streams are heard in a specific sound scene. Bregman (1990) differentiates between an acoustic source as the physical system that generates a characteristic sound pattern, and an auditory stream. The latter denotes the percept which correlated with the said sound pattern in the perceptual world of the listener (Cooke and Ellis 2001).

It is the goal of scene analysis is to recover meaningful descriptions of each separate sound source in the environment. This requires a detailed knowledge of the relations between the physical world and the perceptual (auditory) world of the listener (Bregman 1990). It is likely, that the experienced stream relates to an actual physical object. However, a stream can also form in the case that there is no actual physical object but a conglomerate of sounds which the listeners' mind attempts to organize (Cooke and Ellis 2001), indicating that the auditory system has a tendency to associate each sound property with a certain physical object.

As mentioned above, the Gestalt Rules, which refer to object formation, have originally been formulated for the visual domain. In this context it was considered that the smallest units of perception are rather structured entities than atomistic elements such as single sensory inputs. One of the guiding questions was, why out of a multitude of possible percepts typically only one materializes (Binder et al. 2009), and how the processing of certain cues gives rise to a specific kind of integrative perception, namely, the so-called Gestalt perception. Thereby Gestalt perception apparently happens in a pre-attentive manner.

Wertheimer (1923) was the first to formulate Gestalt rules using the terms *proximity*, *similarity*, *closure*, and *common fate* as principal factors. Under *proximity* he understood that the nearer the stimuli are to one another the more likely they will be assigned to the same object. The same is assumed for *similarity*, meaning that elements with common features such as size, shape, distance, or color, tend to be grouped. *Closure* describes the tendency of the human perceptual system to complete objects, causing closed entities to be formed. Furthermore, stimuli that move simultaneously, for instance, in the same direction or at the same rate, tend to be perceived as a group, as they share a *common fate*. These Gestalt factors were considered mechanisms which are built into the sensory systems through the general evolution of organisms during interaction with the environment, rather than as mechanisms that humans acquire individually. Interestingly, these principles also seem to be effective in a variety of other species, considering that camouflage or concealment mechanisms counteract the Gestalt principles.

The later formulated *Prägnanz* principle targets the question of why a certain interpretation of a given scenery is preferred over other interpretations which are physically possible and plausible as well. *Prägnanz* is thought of as a higher-order concept compared to the principles as described by the Gestalt rules, but may include

proximity, similarity, closure, and common fate. Prägnanz may roughly be described as the tendency of a perceptual scene to become organized in the simplest most homogeneous way possible (Binder et al. 2009) with a preference for concise representations that convey a large informational content while maintaining a high degree of simplicity.

However, the statement that the percept (i.e. the scene) that is perceived is the one with the highest degree of *Prägnanz* can only be helpful for explaining perceptual organization if a precise measure of *Prägnanz* is defined (cf. Jäkel et al. 2016). Two such measures that have been suggested in information theory as measures related to *simplicity* (Jäkel et al. 2016) are *redundancy* and *code length*. Some recent approaches are based on Bayesian inference, which allows a formulation of the likelihood principle in perception and suggests to choose the interpretation which is most likely to be true. Based on the observed data (i.e. the input signals) several interpretations may be possible, each having an associated likelihood function. The subjective belief in a certain interpretation is then associated with its posterior probability—i.e. the conditional probability of the hypothesis given the observed image. It has been shown that under certain assumptions, maximizing the Bayesian posterior is consistent with maximizing the simplicity (Feldman 2009). Froyen et al. (2015) present Bayesian hierarchical grouping, a framework for perceptual grouping based on the assumption that the configuration of the elements of a scene is generated by a mixture of objects that can be described by a mixture model. The challenge of perceptual grouping lies in estimating the generating sources. This causes a conflict between, on the one hand, decomposing the scene into a large number of groups which allow a good fit of the scene data and, on the other hand, decomposing the scene into fewer groups. This would increase the simplicity but does not fit the scene data at the same time. Froyen et al. (2015) suggest to apply Bayes' rule to optimize this trade-off and allow the best balance between simple grouping and a reasonable fit to the scene data.

With such approaches, it becomes theoretically possible to quantify the plausibility of grouping interpretations (Froyen et al. 2015) and thus the degree of *Prägnanz*. Therefore, the elusive notion of *Prägnanz* may eventually become amenable to precise measurement, even though it is still in question how the perceptual system actually determines the relative likelihoods of different interpretations (Feldman 2009).

In modern neuroscience, the Gestalt principles still exist as concepts in the spatio-temporal processing of perceptual features, thus motivating the search for neural correlates of the Gestalt phenomena. One example is the investigation of synchronized coupling of neuronal activity within and across cortical areas that may be interpreted as a correlate of the *common fate* principle (Binder et al. 2009). As the stream plays the same role in audition as the object in the visual domain, the Gestalt Rules can help to solve the auditory-scene-analysis problem. There are a few analogies that can be drawn. For example, the principle of *closure*, which is concerned with completing forms, closing gaps, etc., is useful when a visual object is partly occluded and is likewise necessary when an acoustic signal is masked by an interfering noise (Bregman 1990) (Fig. 2).

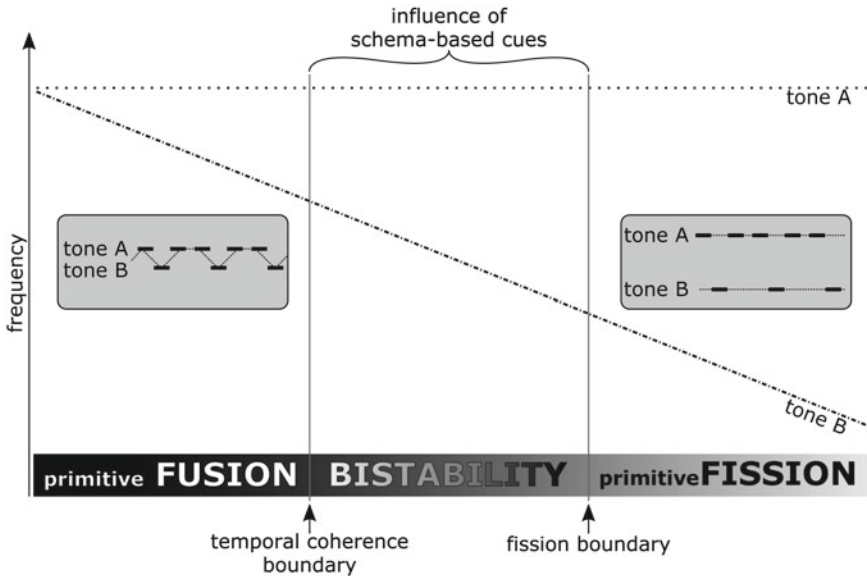


Fig. 2 Schematic illustration of the three states of stream perception as a function of frequency separation between two sinusoidal tones, **A** and **B**. In the underlying experiment as described by Bregman (1990), repeating cycles of two sinusoidal tones (gray boxes) were used to illustrate the importance of speed and frequency separation for the formation of streams. Small frequency separations between **A** and **B** (see left side of the diagram) lead to the perception of one coherent stream, in this case a galloping rhythm, thus characterizing the states of *primitive fusion* or *obligatory fusion*, respectively. Sufficiently large frequency separations will lead to the perception of two separate streams as shown on the far right side of the diagram, marking the states of *primitive fission* or *obligatory fission*. Between both states, that is for frequency separations larger than the *temporal-coherence boundary* and smaller than the *fission boundary*, the perception can switch between one and two streams. In this bistable state, the percept is influenced by various factors, such as exposure time or schema-based cues, like expectations

There are three conditions that roughly describe the state of stream perception. These conditions can be illustrated with the effect of frequency separation between two alternating tones on streaming. For small frequency separations, the listener perceives the tone sequence as one coherent stream, characterizing the state of *fusion*. As the frequency separation is increased, the perception of one coherent stream is changed to two. The threshold at which the segregation appears is referred to as the *fission boundary*. Van Noorden (1975) already used the term *temporal coherence*, in order to describe the impression that the perceived relation between successive tones in a sequence forms a whole, which is ordered in time.

In the converging region of *fission* and *fusion*, the percept can alternate between hearing one or multiple streams. Accordingly, the state in this ambiguity region is termed *bistability* (Moore and Gockel 2012). Usually, listeners report not to hear both percepts at the same time. However, one counter-example is documented by Bendixen et al. (2010), who found that in some cases the listeners hear both an

integrated stream of high and low tones along with a separate stream consisting of only high or low tones.

Within the ambiguity region, different factors influence whether *fission* or *fusion* occurs. One factor is the stimulus duration, as segregation of multiple streams tends to build up over time. When a long sequence of sounds with intermediate frequency separation is presented, the tendency to hear *fission* increases with exposure time (Bregman 1990; Anstis and Saida 1985). A possible explanation is that the auditory system begins with the assumption of only one single source, and fission is perceived only after sufficient evidence was accumulated to support multiple sources. A further influence on the occurrence of segregation is the kind of task that is given to the listener, indicating that it plays a role whether the listener attempts to hear segregated streams or not (Moore and Gockel 2012). In contrast to that, the fission that occurs even when the listener is trying to hear fusion is also called *primitive stream segregation* (Bregman 1990; Cusack and Roberts 2000) or *obligatory segregation* (Vliegen et al. 1999).

When quantifying streaming phenomena, the transitions between the three states are typically observed in dependency on variations of some physical property. The tasks used for measuring streaming usually ask for subjective responses, meaning the participants are asked to classify their perception into different categories—e.g., as either several streams or just one. To control the effects of top-down mechanisms and aim for the boundary to *obligatory fusion* or *fission*, the successful completion of the task should depend on achieving either of the two states, for example, by requiring the recognition of two interleaved melodies or recognizing a certain cue which can only be done if the listener is able to segregate streams (compare Dowling 1973; Hartmann and Johnson 1991; Moore and Gockel 2012). Inferred from such experiments are factors and feature types that promote or inhibit segregation, respectively shifting the thresholds between the streaming states to different parameter values.

2.2 *Decomposition of Acoustic Scenes with Binary Masks*

When entering the ear and passing auditory periphery, the sound waves undergo different transformations during which the signal is decomposed into frequency bands and eventually transduced as a pattern of nerve-firings. There is reason to believe that the auditory periphery provides the brain with an internal representation that is similar to a temporal sequence of short-term spectrograms. An auditory stream as described in the previous section could then be regarded as the perceptual grouping of the parts of this neural spectrogram that belong together (see Bregman 1990; Cooke and Ellis 2001). In a similar manner, the CASA systems typically transform the input signal into a time-frequency representation approximating the behavior of the auditory periphery. An essential part of this transformation is the cochlear filtering which is commonly modeled with a filter-bank or a cochlear model. The filter-bank mimics the frequency-selective behavior of the basilar membrane by simulating the frequency response of a point along the basilar membrane as a filter output. A typical

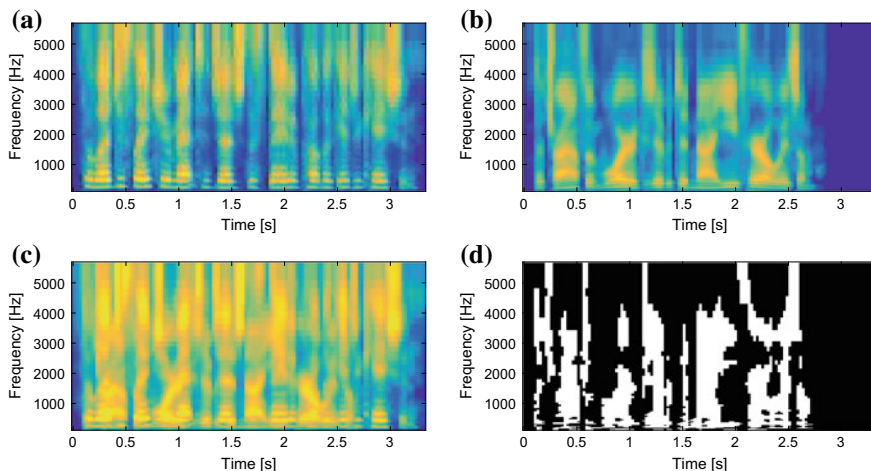


Fig. 3 Cochleagrams and ideal binary mask of two single speakers and the mixture of both speakers. The cochleagrams are generated by passing the time signals through a Gammatone filter-bank with center frequencies ranging from 100Hz to 6kHz that are equally spaced on the Equivalent Rectangular Bandwidth (ERB) scale which is a perceptually motivated division of the frequency axis (Glasberg and Moore 1990). The filter outputs are divided into 50ms time frames with 50% overlap. **a** Cochleagram of the female speaker uttering the phrase “The legislature met to judge the state of public education”. **b** Cochleagram of the male speaker uttering “The triumphant warrior exhibited naive heroism”. **c** Cochleagram of the speaker mixture. **d** An ideal binary mask indicating the time-frequency units dominated by the male speaker—**white pixels**

output is presented in Fig. 3. Here, the time signals are passed through a Gammatone filter-bank with center frequencies that are uniformly spaced on the equivalent rectangular bandwidth (ERB) rate scale (Glasberg and Moore 1990). The filter-bank output is then divided into time frames of 50ms length. By calculating the power within each time-frequency unit, a cochleagram is generated as presented in panels (a) to (c) of Fig. 3.

It should be considered that phase delays can be introduced by the filter-bank, and for some applications, such as the comparison of onset and offset times, compensation of the phase delay becomes necessary (Brown and Cooke 1994). Another option to further process the filter-bank outputs is a model of auditory-nerve transduction such as the Meddis model. It can be used to generate a representation of the firing rate of a nerve fiber by simulating processes in the auditory nerve such as rectification, saturation, and phase locking. The model computes the probability of spikes in the auditory nerve (Meddis 1988).

Adapting the concept of a stream, the time-frequency units dominated by the same source will form an auditory object. The aim of most CASA systems is to identify the ideal binary mask (IBM) which labels the time-frequency units associated with the target source. Consequently, after transforming the signal into a time-frequency representation, each element of the representation is labeled “1” if dominated by the target energy and “0” else. Narayanan and Wang (2013a) define the IBM as follows.

$$\text{IBM}(t, f) = \begin{cases} 1, & \text{if } \text{SNR}(t, f) > \text{LC} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where t and f label time frame and frequency channel. LC refers to the local SNR criterion, above which a time-frequency unit is set to 1. The idea of this mask is to retain the fragments where the target energy is relatively high and remove all other fragments.

The IBM was originally inspired by the auditory masking phenomenon (Wang and Brown 2006). Units that are labeled “1” correspond to unmasked time-frequency units, whereas the “0”-labeled units correspond to masked fragments (Narayanan and Wang 2013a; Li and Wang 2008). The concept is supported by studies which show that within a critical band, the weaker signal is masked by the stronger one (Moore 2012). An argument for the sufficiency of IBMs for the reconstruction of speech can also be drawn from studies on *glimpsing*. The idea of *glimpsing* is that the essential information for human speech perception is conveyed by the time-frequency units with a favorable local SNR and, as such, they offer a *glimpse* of the target (Cooke 2006). For a mixture of two speech signals, the regions of high energy are sparsely distributed, that is, there is relatively little energetic masking in terms of spectrotemporal overlap, and a large proportion of target speech is little affected by the interfering speaker. Studies of speech intelligibility after spectral filtering (Warren et al. 1995; Lippmann 1996) or with severely degraded information regarding the distribution of spectral energy (Drullmann 1995; Shannon et al. 1995) suggest, that much information contained in a speech signal is redundant for the task of speech recognition. This redundancy allows speech to be identified based on relatively sparse evidence. Cooke (2006) applied ASR to glimpses of target speech to identify consonants in various masking conditions that offered differing glimpse sizes. Outputs of this computational model were compared to listeners’ performance on the same task (Simpson and Cooke 2005). The results suggest that the amount of glimpses is a good predictor for speech intelligibility and not the global SNR on its own (Cooke 2006). Furthermore, the recognition scores confirmed that the glimpses are sufficient to support consonant identification.

Target speech which is reconstructed from IBMs can substantially improve the intelligibility in different masking or reverberant conditions as compared to not using any IBM prior to the reconstruction (Roman et al. 2003; Brungart et al. 2006; Li and Loizou 2008; Roman and Woodruff 2011). A similar benefit from the application of an IBM can be observed for the performance of automatic speech recognition (ASR) (Srinivasan et al. 2006; Narayanan and Wang 2013a).

The construction of the IBM requires access to the cochleagrams of the premixed signals in order to calculate the local SNR. Since these premixed signals are unknown in most real-life scenarios, the computational goal of CASA is to estimate the IBM *blindly*, that is, based only on observable physical features and without information that would usually not be available to the listeners (Wang and Brown 2006). The underlying assumption is that the dominant physical features that are available in the given time-frequency representation provide enough evidence to determine the locally dominant source.

Once a binary mask has been estimated, it can be used for the reconstruction of a single speaker or speech signal. In ASR, the masks have been mainly used in missing-data frameworks (Cooke and Ellis 2001; Raj et al. 2004). Some systems have applied the binary masks to either marginalize the probability of masked features or to reconstruct them by making use of prior distributions of speech and the reliable unmasked features (Raj et al. 2004). It has also been shown that binary-masked signals can be used based on missing-data methods without marginalization or reconstruction steps (Hartmann and Fosler-Lussier 2011; Hartmann et al. 2013; Narayanan and Wang 2013b). Improvement of overall ASR performance and intelligibility can further be gained by replacing the zeros in the binary mask with a low noise floor prior to resynthesis (Cao et al. 2011).

Despite the observed benefits of IBMs, it is worthwhile to consider alternatives to this approach. One required property of the IBM is that it is supposed to produce an output with the highest global SNR gain among all binary masks, namely, the SNR gain averaged across the entire time-frequency plane. Li and Wang (2008) address the global optimality of IBMs in terms of signal-to-noise ratio. They show that despite the locally optimal SNR gain, the IBM is not necessarily globally optimal in the same sense unless the time-frequency decomposition is orthonormal, which is not the case for many types of commonly used time decompositions with overlapping frames. Thus, the optimality in terms of global SNR gain is not generally given. Another weakness is that non-dominant components of the target are neglected.

Phenomena such as the binaural masking-level-difference or the equalization-cancellation model describe human listening phenomena that rely on non-dominant physical features which are inevitably lost after the application of a binary mask, because they appear in time-frequency units in which the target signal is non-dominant. Furthermore, it seems that listeners make use of glimpses with moderately negative local SNRs. Cooke (2006) and Narayanan and Wang (2013b) found that the commonly used local SNR criterion of 0dB does not maximize the ASR performance. Rather, ASR word accuracy as well as human speech intelligibility for IBM-processed mixtures of speech and noise showed best performance with a local criterion below 0dB.

Schoenmaker and van de Par (2016) investigated whether negative-SNR target-speech components contribute to speech intelligibility in human listeners. To this end, they locally removed such components from test sentences with multiple simultaneous talkers. They observed that in comparison to a reference condition in which the criterion for removing target speech components was set to $-\infty$ dB SNR, only the condition with -4 dB SNR showed no significant difference in speech recognition rate. Higher criteria led to a decrease in speech intelligibility, indicating that essential information for speech recognition is still present in regions with slightly negative SNR, that is, above -4 dB. A conservative IBM estimation relying only on the dominant features would probably discard these regions.

An alternative is the ideal-ratio mask (IRM) which considers the ratio of target energy relative to the mixture energy within each time-frequency unit. In an ASR experiment (Wang and Wang 2016) observed better recognition rates using direct recognition of IRMs than using direct recognition of binary masks. Yet, according

to Li and Wang (2008) the SNR gains due to IBMs with benefit due to IRMs and SNR were similar. Estimation of IRMs requires estimation of the energy ratio of two signals while the IBM estimation is facilitated by numerous classification and clustering methods (Li and Wang 2008).

3 Object Formation

The different types of objects in our environment each encompass a characteristic set of sound features, conveying information about size, material, distance and other attributes of the object. Knowledge of these characteristic features lays the foundation for tracing the features back to the object. Some features, such as pitch, offset time, or spatial position, can be directly connected to the physically tractable signal properties and are referred to as *primitive cues*. Other features are strongly influenced by preconditions of the listener, for instance, attention, familiarity with certain sounds, or knowledge of the spoken language. These are referred to as *schema-based cues*.

In the following section, both principles will be presented, along with features that underlie object formation in the human auditory system. Firstly, primitive cues are dealt with by looking at findings from ASA, the physical background, limits of feature detection, and implementation methods. Secondly, schema-based cues are considered, that is, features that are not inferred from the physical stimulus alone but rely heavily on previous knowledge and attentive mechanisms. Examples of both feature types are presented along with some concepts on how the two processes interact.

3.1 Primitive Grouping

Physics of Sound as the Basis of Grouping Cues

Primitive cues can be extracted directly from the acoustic input signals to the two ears and are assumed to be represented in the periphery of the auditory system. In data-driven approaches to ASA, these cues are treated as the basis for auditory object formation. This type of processing is also referred to as *bottom-up processing*, due to the direction in which information is passed from the lower-level representations to higher cognitive processes. This already implies that interpretations of primitive features are innate and do not rely on prior knowledge.

Primitive cues and acoustic events are linked via the mechanical processes of sound generation and transmission. Alías et al. (2016) suggest a rough classification of sound events into three groups, that is, speech, music and environmental sounds. Speech is produced in the vocal tract with the help of vocal chords and resonant cavities to produce vowels, and with tongue, teeth, and lips to produce noisy

components, such as consonants. The physical consequences of this process are fundamental frequencies within a limited range of frequencies as determined by the vocal chords, a pattern of harmonics shaped by the resonant cavities and relatively smooth transitions of these frequencies over time. In a similar manner, the sound production with musical instruments is bound to an instrument-dependent range of fundamental frequencies and the spectrum of harmonics produced by its resonator. Both speech and music exhibit a high degree of periodicity in the time domain and are composed of a limited dictionary of sound units (phonemes and notes). The category “environmental sounds” comprises the remaining sound events, for instance, noise, nature-based, machine-generated, and human-activity-based sounds. However, there are fewer generalizations or common constraints about the sound properties in this category than for speech and music.

The perceptual effects of physical sound properties have been studied in various psycho-acoustic experiments. These experiments have explored the correlations between acoustic (physical) events and auditory (perceptual) events, eventually allowing to draw conclusions about the grouping information that the auditory system uses. Broadly speaking, the majority of grouping cues relies on or is derived from either onsets and offsets, temporal modulations or spatial features of the acoustic events—(compare Cooke and Ellis 2001).

On- and offsets refer to sudden changes of intensity in the time signal. If the on- or offsets of the acoustic signals in different frequency bands occur at the same time instants or temporally close instants, they are likely originating from the same physical object—the sound source. In a similar fashion, the resemblance of temporal modulations (which includes rhythm and frequency) indicates the same source of origin. Vliegen et al. (1999) observed that fundamental frequency and spectral shape influence the stream segregation. Remarkably enough, the effect of differences in fundamental frequency alone varies as it seems to be stronger whenever segregation would be advantageous but is weaker than spectral differences in producing *obligatory segregation* whenever *fusion* is advantageous (Grimault et al. 2000). Segregation based on the estimated fundamental frequency is also referred to as pitch-based segregation and especially used for the segregation of voiced speech. Also slower rates of temporal modulations that would not be perceived as pitch but rather as a rhythm, for example in the envelope of the signal, affect stream perception and can, in the case of large differences in modulation rates of the envelopes, enhance segregation (see Moore and Gockel 2012, for a review).

Spatial cues include interaural level differences (ILDs), interaural time differences (ITDs), and monaural spectral cues as are mainly introduced by the pinna. In an experiment with a sequence of consonant-vowel tokens, David et al. (2017) showed that listeners use both interaural-difference cues and monaural spectral cues in order to segregate syllable sequences. However, they found that only when all spatial cues (ITDs, ILDs, and spectral cues) were present, obligatory stream segregation (i.e. fission) was produced. Studies on the effect of perceived location on streaming by manipulating ITDs found that the apparent spatial location of the auditory event produced by ITD manipulation alone had a relatively weak effect on obligatory stream segregation (Stainsby et al. 2011; Füllgrabe and Moore 2012). Nevertheless,

differences in the apparent location produced by ITD differences between target and masker still support stream segregation under conditions where segregation is advantageous (Moore and Gockel 2012). A similar indication has been formulated by Schwartz et al. (2012), namely, that the full complement of spatial cues along with harmonicity and common on- and offsets, allows listeners the most accurate segregation of a sound source and identification of its content. Consequently, none of the above cues seem to be redundant despite a case-by-case dominance of certain cues.

An important question with respect to the organization of sounds is the following: Which features are employed for object formation and how are they integrated within the auditory system in the formation process? The information originating from a single source is not only distributed across time and frequency but also represented in different feature types. With regard to neural processes, they are likely to be distributed across different brain centers as well (compare Cooke and Ellis 2001). At this stage, the Gestalt Rules provide some idea on how the disseminated information may be collected and how the brain may form connections between the single elements of the sensory input. The Gestalt Rules mainly refer to an automatic or innate organization associating them strongly with primitive grouping cues. When translating the visual Gestalt Rules into the auditory domain, only *proximity* in space and time are directly transferable. As for the other principles, it is necessary to find auditory counterparts. For example, in addition to proximity in time and space, the proximity in frequency becomes more important in the auditory domain. While in visual terms surface, texture, and color are features suitable for *similarity* estimations, in the auditory domain timbre, amplitude and pitch, are useful features. *Common fate* which strongly refers to spatial attributes or movements in the visual domain, has to be interpreted rather in terms of spectrotemporal patterns. Pitch or harmonics, for example, cause different regions of the spectrogram to be activated at the same time and rate, or show similar amplitude or frequency modulations.

Application of the principles of *similarity* or *proximity* can be seen as grouping of sound components with similar timbres or proximity in time and frequency (Bregman 1990). Since several features vary simultaneously if produced by a single source, it is also sensible to group simultaneously changing features according to the *common fate* principle.

Grouping information appears to be accumulated across time. This is observed in the build-up of stream segregation or fission with increasing stimulus duration. A similar effect is shown by Snyder et al. (2009), who present evidence that even after a stimulus has already stopped, it has a persisting effect on stream segregation, lasting from tens of seconds to seconds. Although this effect starts to decay after a few seconds, it can take many seconds of silence to decay completely (Moore and Gockel 2012). Thus, it seems that a central process is integrating grouping information based on different cue types. Partly, this may be attributed to the possibility that certain cues, such as accurate pitch estimation, require a sufficient signal duration or averaging time before the temporal modulation can be reliably identified or even recognized (Cooke and Ellis 2001). However, it should also be considered that the auditory system puts certain constraints to feature perception, for instance, due to limited

total bandwidth or limited resolution. Higher frequencies remain unresolved as they fall in the relatively broader auditory filters than lower frequencies. Further, the envelope of the output signal of the filters becomes perceptually more important with increasing frequency. Effects of this kind within the auditory system modify the feature perception in combination with the physics of sound generation and, consequently, determine the cue types which can be exploited for ASA.

Limits of Feature Detection

The detection of features that are present in a complex signal is a process that is inherently limited due to uncertainties about the exact properties of the features and their distortion caused by the presence of interfering sources. The theoretical limit in the accuracy with which features can be extracted is often difficult to determine. It is, however, possible to examine the most simple feature extraction that can probably be considered, that is, detecting a known signal in the presence of a white noise interferer and reflect the theoretical limitations in the accuracy of this feature. Feature detection would in principle only entail a “yes” or “no” decision regarding the presence of the known signal.

When detecting a known signal within white noise that is sampled at a certain rate, it can be derived that the most optimal manner to detect the signal is to create a so-called *matched filter*. This filter works on the assumption that the signal-to-be-detected itself is buried in white noise and provides an optimal filter to detect it in white noise. Yet, even with this detector performing optimally, its performance is limited by the number of samples that represent the number of independent observations and by the signal-to-noise ratio. The more samples that are available, that is, the more degrees of freedom in the noise signal, and the higher the signal-to-noise ratio, the better the matched filter will perform.

If the bandwidth of the noise is limited, the assumption of independent sampling is violated due to autocorrelation within the signal. This results in poorer performance. Down-sampling and spectral whitening, such that the represented bandwidth coincides with the noise bandwidth, would again ensure that samples are independent. Nevertheless, it has to be concluded that for narrowband noises, optimal signal detection yields poorer performance than for broadband noise. This notion of a fundamentally-limited performance, even when the signal-to-be-detected is exactly known, extends to general feature extraction. Typically, it can be expected that feature extraction will become noisier, the shorter in time or the narrower in bandwidth the signals are from which the features are extracted—in other words, the fewer degrees of freedom there are in the signal.

This fundamental limitation in feature extraction, which requires the extraction to be based on larger spectrotemporal intervals, provides a fundamental challenge in devising CASA systems. On the one hand, reliable feature extraction requires larger spectrotemporal regions to be used for the extraction, while at the same time, possibly fine decisions need to be made about specific spectrotemporal regions belonging to a particular source.

Implementing Primitive Grouping in Algorithms

The formation of auditory objects under unknown conditions requires the exploitation of intrinsic physical cues that allow a bottom-up type of processing. Thereby, the initial build-up of a new auditory scene relies particularly on primitive cues. Computational challenges concerning the implementation of primitive grouping lie in the extraction of suitable cues, an estimate of their reliability, and the formation of plausible connections between different types of evidence.

There exist a number of computational models that attempt to explain the use of primitive grouping cues in human hearing. Beauvois and Meddis (1991, 1996) present an early model to simulate aspects of stream perception for tone sequences with alternating frequency. By manipulating the repetition time and frequency separation, they were able to explain grouping through proximity in frequency and time as well as the temporal build-up of streaming, as had been shown earlier by Anstis and Saida (1985). A model for vowel segregation by Meddis (1988) applies auto-correlogram analysis to identify frequency channels which respond to the same voice. This model was able to automatically recognize vowels with a performance close to the results of listening tests with humans (Assmann and Summerfield 1994; Cooke and Ellis 2001). CASA systems are not necessarily geared at being physiologically plausible but often apply similar computational methods as physiologically motivated models.

A common method to retrieve pitch information is the analysis of the autocorrelation function. An example for calculating the normalized autocorrelation (NAC) from a time-frequency representation is given as follows.

$$\text{NAC}(t, f, \tau) = \frac{\sum_{n=0}^{N-1-\tau} [x(t, f, n) \cdot x(t, f, n + \tau)]}{\sqrt{\sum_{n=0}^{N-1-\tau} x(t, f, n)^2} \sqrt{\sum_{n=0}^{N-1-\tau} x(t, f, n + \tau)^2}}. \quad (2)$$

where t and f are the time frame and the subband index. τ denotes the time lag in samples.

Chen and Hohmann (2015) used this NAC calculation for a pitch estimation by combining it with the calculation of the comb-filter ratio between time-frequency units. In their algorithm, time lags in a range of 2.4–14.3 ms were analyzed, which correspond to fundamental frequencies between 70 and 420 Hz. By subband averaging the most salient pitch within one time frame is obtained. Cross-channel correlation of the normalized autocorrelation response can also be used to estimate the pitch strength in a frequency channel (Wang and Brown 2006). With L being the maximum time lag of the correlogram in sampling steps and the normalized autocorrelation of the filter output, $\hat{A}(f, t, \tau)$, the cross-channel correlation can be calculated according to the following equation.

$$C(f, t) = \frac{1}{L} \sum_{\tau=0}^{L-1} \hat{A}(f, t, \tau) \cdot \hat{A}(f + 1, t, \tau). \quad (3)$$

Under the assumption that neighboring and partly overlapping frequency channels respond to the same frequency component their correlation should be relatively high if they are dominated by a periodic signal.

A representative system that makes use of pitch cues for speech segregation in reverberant signal mixtures is presented by Roman and Wang (2006). Prior to the pitch-based segregation, a filter is applied to the reverberant mixture which inverts the impulse response of the target room. As a result of this stage, the periodicity of the target is enhanced while the signals arriving from other directions are further smeared. For the pitch-based segregation, a correlogram is computed based on the time-frequency representation. To generate the correlogram, the autocorrelation is computed at the output of each frequency channel. In high-frequency channels, the envelope of the filter response is regarded instead of the fine structure of the time signal. The extracted periodicities are compared with an estimated target pitch and grouped if the underlying target is stronger than the interference. Eventually, the segments likely to originate from the target are grouped in a binary mask. Hu and Wang (2010) present an algorithm that jointly and iteratively performs the pitch estimation of the target and the segregation of voiced portions. Periodicity is indicated by peaks in the corresponding autocorrelation function, also considering neighbouring time-frequency units to reduce errors. After an initial estimation of the target pitch, the estimate is used to segregate target speech using harmonicity and temporal continuity. A time-frequency unit is labeled “1”, if it exhibits a periodicity similar to that of the target.

On- and offsets are specifically interesting for segregating components that are not captured with periodicity analysis. Wang and Hu (2006) and Hu and Wang (2007) suggest on- and offset analysis for the segregation of unvoiced speech. To detect sudden intensity changes that correspond to the on- and offsets, the intensity of each filter output is smoothed at a different degree. The higher the degree, the smoother the output. This reduces random intensity fluctuations. The first-order derivatives of the smoothed outputs are then calculated and the peaks or valleys are marked assuming that these represent on- and offsets. By matching close peaks or valleys, the system forms on- and offset fronts which are then used to assemble larger segments. By considering different degrees of smoothing, the issue of under- or over-segmentation is handled. Under- or over-segmentation is generally caused by a too sensitive or too coarse on- and offset detection.

An exemplary implementation of binaural cues for scene analysis was suggested by May et al. (2011). ITDs and ILDs are jointly analyzed to determine the azimuth position of the source. The ITDs are calculated for each channel, f , and time frame, t , using the normalized cross-correlation between the ears, C , for time lags in a range of $-1, 1$ ms.

$$C_f(t, \tau) = \frac{\sum_{n=0}^{N-1} (l_f(t \cdot \frac{N}{2} - n) - \bar{l}_f) (r_f(t \cdot \frac{N}{2} - n - \tau) - \bar{r}_f)}{\sqrt{\sum_{n=0}^{N-1} (l_f(t \cdot \frac{N}{2} - n) - \bar{l}_f)^2} \sqrt{\sum_{n=0}^{N-1} (r_f(t \cdot \frac{N}{2} - n - \tau) - \bar{r}_f)^2}}, \quad (4)$$

where \bar{l}_f and \bar{r}_f denote the mean values of the left and right auditory signals that are estimated over the time frame, t . The time lag for which the cross-correlation function exhibits a peak corresponds to the estimated ITD (in samples). To obtain ILDs, the energy in each ear was integrated across a time interval, N , and compared between left and right ear (expressed in dB).

$$\text{ILD}_f(t) = 20 \log_{10} \left(\frac{\sum_{n=0}^{N-1} r_f \left(t \cdot \frac{N}{2} - n \right)^2}{\sum_{n=0}^{N-1} l_f \left(t \cdot \frac{N}{2} - n \right)^2} \right). \quad (5)$$

The ITDs and ILDs are fed into pre-trained Gaussian-Mixture Models, which convert the binaural cues to probabilities for different azimuth positions.

One of the intricate aspects in CASA is the combination of different feature types. Brown and Cooke (1994) suggested an early CASA system which uses a combination of periodicity, frequency transitions and on- and offsets in auditory nerve firing patterns for the segregation. The different feature types are represented in separate auditory maps which serve as intermediate representations between acoustic input and a symbolic description of the input. In these time-frequency maps, the values of the regarded feature type are represented on an orthogonal axis, such as in an autocorrelation map, a frequency transition map, and an onset map. Elements which simultaneously change in a similar way, for instance, common fundamental frequency or on- and offset times, are grouped—similar to the *common-fate principle*. A combination of monaural and binaural cue analysis is presented by Woodruff and Wang (2010). For the simultaneous organization across frequency and short continuous time intervals, monaural cues are used. The segments obtained from this step are then sequentially organized by regarding the averaged binaural localization cues. In the system presented in Woodruff and Wang (2013) pitch and localization cues are jointly analyzed and both used for simultaneous organization.

Often, the use of primitive grouping cues is to some degree combined with previously trained statistical models such as Gaussian-Mixture Models in which ITDs and ILDs are converted to azimuth positions. The localization model of Josupeit et al. (2016) extracts instantaneous ITD information, which is then combined with a type of schema-based organization by using a template-matching procedure based on periodicity and spectral energy to select target-related ITD information. Mandel et al. (2010) separate and localize sound sources based on interaural phase and level differences. The system models each source probabilistically based on their interaural parameters and evaluates each point in a spectrogram to identify regions which best-fit each of the respective source models.

3.2 Schema-Based Organization

Although very important, primitive cues typically only group signal components on a very local basis in the time-frequency plane. To link these local structures into more global auditory objects, knowledge about the behavior of these objects must

be incorporated into the process of ASA, for example, to identify moving sources, or sources that change in timbre. In the following, it is discussed how the human auditory system builds its rules for streaming, and how CASA streaming is performed by applying rules that are encoded explicitly or learned from training data.

Schema-Based Processing in Human Listening

Cues for auditory organization that involve learned rules and attentional mechanisms are referred to as schema-based features. Complementary to the *bottom-up cues* these are also considered *top-down cues*, since high-level representations induce information used to form the auditory object. These *schemata* that supplement the primitive cues are assumed to be learned patterns of speech, music or environmental sounds. Such schemata can add robustness to the auditory scene analysis since the listeners exploit the familiarity with language or other sources to fill masked or distorted parts of the signal. This enables the listeners to deal with very limited information. One example for this completion is the *phonemic restoration* effect. Listeners that were presented with sentences were unaware that short segments of the sentences were removed when replaced by a louder noise like a cough. Instead of detecting the gap, the speech was heard as complete (Warren et al. 1972). In retrospective, the listeners were not able to specify the exact timing of the cough or distinguish directly-heard speech sounds from restored ones. The phenomenon demonstrates how the auditory system actively engages in the interpretation of inputs and creates a structure or even an illusion based on the signal context—(see Cooke and Ellis 2001). Linking this observation to the previously described Gestalt Rule of *closure*, it may be concluded that both in the visual and the auditory domain, the human perceptual system has a tendency to complete objects.

It appears that the time span across which schema-based processes operate is longer than the span over which primitive grouping cues are important and that, while local spectrotemporal cues influence the object formation on a syllable level, the schemata are key to auditory stream formation (Bregman 1990; Shinn-Cunningham et al. 2017). Kidd et al. (2014) studied the influence of syntax on speech identification in masked speech and found that the listeners' performance was significantly better in cases of correct target-sentence syntax than with incorrect syntax. This indicates that the predictability of elements through syntax supports the formation and maintaining of streams.

Another powerful strategy to efficiently deal with the sparse information is the integration of constraints concerning what is known to be a plausible interpretation of the sparse input. These restrictions of possible interpretations are considered to be caused by *expectations*, that is, a bias towards a certain interpretation. The effect of expectations becomes particularly evident when a source is perceived due to expectations built up by the signal context, even in cases when physical cues of the source are not present—(see Cooke and Ellis 2001, for a review). Besides the previously described phonemic restoration, this can be illustrated by the *continuity illusion*. A tone that is briefly masked by a noise burst is typically perceived as continuing during the noise, despite the absence of physical cues during the masked period.

Strong bias toward such expectations seems to be the assumption of speech, causing the auditory system to presume that signals with any speech-like character are indeed speech. Experiments with sine-wave speech (Remez et al. 1981), which lacks the traditional speech cues, demonstrated that relatively little information, such as the time-varying properties of speech, is sufficient to perceive the linguistic message although the listeners judged the quality of the voice as unnatural. Hence, it seems that the existence or absence of local cues can be invalidated if they are combined with linguistic or other expectation-related constraints. As Cooke and Ellis (2001) describe it, the perception exists as a compromise between direct physical evidence of sources and the absence of contradictory cues.

Another aspect to be considered when regarding top-down mechanisms is voluntary *attention*. While primitive grouping is assumed to operate rather automatically and pre-attentive, schema-based organization is assumed to partly involve voluntary attention. Attention can be consciously experienced as the selection of an auditory object for further analysis or can be involuntary, for example in situations when attention is unintentionally drawn to a certain event. The extent to which attention is necessary in order to build-up the tendency to hear stream segregation is not yet defined (see Moore and Gockel 2012 for a review) but different studies (Thompson et al. 2011; Carlyon et al. 2001, 2003) indicate that the build-up of stream segregation is reduced when the attention to a sequence is absent or switched between ears.

Masking as attributed to these higher-level processes is commonly summarized under the term *informational masking* complementary to *energetic masking*, the latter being related to spectrotemporal overlap (Durlach et al. 2003). Studies on energetic masking give an insight on the extent by which top-down processing enhances or inhibits the organization of primitive cues. Arbogast et al. (2002) investigated the effect of spatial separation on masking of speech for different types of maskers that ranged from energetic to informational. Their results showed that speaker segregation benefited most from the spatial separation in the condition with primarily informational masking as compared to primarily energetic masking. Specifically interesting for multi-talker situations are also the voice characteristics and the number of talkers. Brungart et al. (2009) investigated the effects of both factors in multi-talker situations and observed the portion of energetic masking by applying ideal time-frequency segregation. Their results show that *energetic masking* increased systematically with the number of competing talkers while the target-masker similarity had a small systematic impact on energetic masking. Their results suggest that non-energetic masking due to a confusion of target and masking voices are assumed to play a more significant role in this case. With regard to the implementation in CASA and the estimation of the IBM, which reflects the amount of energetic masking, the IBM will become smaller if energetic masking increases. As informational masking comes into play, the identification of the ideal mask is aggravated. This is a challenge for both the human listeners and machine-listening.

Knowledge-Driven Algorithms

In CASA, schema-based processing is generally implemented by specifying object behavior by a combination of explicitly stated rules—curated by researchers and developers—and parameters learned from carefully selected and processed training data. In the case of speech signals, this could be in the form of N-Grams or Hidden Markov Models (HMM), which are based on the idea that having observed a given class of signals, for instance, specific speech sounds (phones), it is possible to estimate the probability of observing the subsequent state of the model. These models can be scaled to higher structures accounting for linguistic structures, such as the level of sentences or paragraphs. While on the phoneme level the HMMs are trained using large corpora of speech signals, for higher-level structure analysis it is not uncommon to use explicitly-coded linguistic rules.

The framework FADE (simulation framework for auditory discrimination experiments) (Schädler et al. 2016) uses an HMM-based automatic speech recognizer to predict the outcome of auditory experiments such as speech intelligibility. In this framework, the HMMs are used to model a number of states for each word of a sentence test, supplemented by models of start, stop, pre-silence, and post-silence. Further, a word network (that is, a representation of possible word successions) to account for the grammatical structure of the test. These elements can be regarded as an implementation of the knowledge of a trained listener who is familiar with the limited vocabulary of the test as well as its syntax. Evaluations of FADE using different feature types as input to the HMM training and recognizer showed that a single set of general parameters can be found which allows the simulation of a variety of different experiments. Furthermore, it was observed that while in some of the tested noise conditions the simulated Speech Reception Thresholds (SRTs) were dependent on the type of input features, in other conditions the SRTs were not dependent on the difference in the tested feature types. A similar approach was taken by Spille et al. (2017) who successfully predicted human SRTs using an ASR-system that implements a deep neural network (DNN) to convert the acoustic input into phoneme predictions and thereby allowed speech-intelligibility prediction of unseen speech signals.

In the domain of music-source separation, another form of schema-based processing can be found in the form of non-negative matrix factorization (NMF). Here, the model of sources is represented by codebooks that are learned from training data (e.g., the spectra of all the sounds an instrument is capable of producing Smaragdís and Brown 2003). The designer can encode specific rules (e.g., restricting the codebook to the notes of a western scale with a set range), all of which are then explicitly learned in training. The schema can then be extended to higher levels using multi-level NMF (Ozerov et al. 2011), where recurring temporal activations of codebook elements can be encoded in higher-level dictionaries.

3.3 *Interaction of Bottom-Up and Top-Down Processes*

The interplay between primitive and schema-based mechanisms needs to be considered when investigating the process of object formation in the auditory system. It seems that while the primitive processes are mainly concerned with partitioning the sensory input, the schema-based mechanisms select from the input (Bregman 1990). Shinn-Cunningham et al. (2017) suggest that the two processes of object formation and selection are what enables humans to deal with the Cocktail-Party Problem, that is, understanding a specific talker in an auditory scene situation with multiple simultaneously active speech sources. Their approach is to not view this as a hierarchical process in which objects are first formed and thereafter selected for further analysis. Instead, the course of object formation should be treated as a heterarchical process during which formation and selection influence each other. From a neuroscientific perspective, it is not yet possible to determine the neural processing stages at which object formation occurs, but it is unlikely that there is one specific stage at which the object first appears. Rather, an object-based representation seems to gradually develop along the auditory path while attentional selection can occur at every stage.

Research on object formation in the visual domain indicates that in the periphery of the visual system, the representation is strongly determined by the pattern of the entering light and less affected by what the perceiver is trying to process. However, it seems that the influence of attention increases at each progressive processing stage while the influence of the actual light input decreases. The common cases where listeners fail to hear a sound component that is well represented in the auditory nerve, support the role of central limitations on both detection and recognition (Shinn-Cunningham et al. 2017). Their ability to override bottom-up ASA mechanisms indicates that the segregation stage cannot merely be seen as the preliminary stage for recognition, but that there seem to be mutual influences. Investigating the role of top-down mechanisms in streaming, (Bey and MacAdams 2002) presented listeners with two unfamiliar melodies, of which one was interleaved with distractor tones. The listeners' task was to indicate whether the two melodies were the same or different. In one condition, the undistracted melody was presented first while in the other condition the order was switched. They found, provided a sufficient difference of mean frequency between target melody and distractor, that the performance was better in the first condition. They interpreted this as an argument that schema-based mechanisms can only operate after a certain amount of primitive segregation has already taken place. The interpretation was questioned by Devergie et al. (2010), who applied a similar task with interleaved melodies but used highly familiar melodies. The results showed that the performance was still above chance when the target and distractor melodies were in the same frequency range and concluded that schema-based mechanisms can indeed function without preliminary primitive segregation, provided that the melodies are highly familiar (Devergie et al. 2010). Hence, it seems that the primitive segregation is of higher importance whenever the input is unfamiliar.

The difficulty of putting primitive and schema-driven grouping into a fixed order becomes obvious in Cocktail-Party situations. On one hand, there is evidence that listeners are not actually able to divide attention between the different sources. A certain degree of *change deafness* is observable depending on which auditory object they are attending to, displaying a dominance of schema-driven processing. On the other hand, certain primitive factors, referring to the statistical salience of the auditory stimuli, such as unexpectedness and uniqueness, show that bottom-up mechanisms can as well override attentional top-down processes—(compare Shinn-Cunningham et al. 2017).

4 Conclusion

The intention of CASA algorithms is to reliably extract auditory objects from sound mixtures, drawing inspiration from principles of human hearing. Current systems are able to handle different feature types and deal with noisy or reverberant environments. However, the human auditory system still exhibits more flexibility and robustness than current CASA algorithms when it comes to unfamiliar listening conditions with unknown types or numbers of sources. Some technical implementations of CASA principles have been presented in this chapter, yet many aspects remain that are fairly complex and have not yet been realized in algorithms, for example, the modeling of interactions between top-down and bottom-up mechanisms and the integration of attentional mechanisms. Also, the estimation of binary masks relies on the extraction of dominant features, while experiments show that both human listeners and ASR systems benefit from moderately negative local SNRs.

An issue which is treated more commonly is the integration of different feature types. Systems that rely on one feature type are prone to instability, not only because the specific feature type could be masked but also due to the ambiguity of single cues, which could be solved through integrating further evidence. There are results from perceptual studies which indicate that listeners make use of certain cues to compensate for changes in other cues, suggesting that at some level all cues are mapped to a single perceptual attribute—(compare Cooke and Ellis 2001). In Sutojo et al. (2017) a framework is presented that targets this issue but remains to be extended in future work. The main idea is to combine different features to derive a similarity value for neighboring time-frequency units and thereby obtain a single grouping attribute. The weights with which each feature influences the similarity value are obtained through prior training. Based on the estimated similarity value, a grouping decision between direct neighbors is made. With regard to the auditory Gestalt rules, this approach mainly exploits the principle of similarity according to which the elements (which in this case are time-frequency units) with common auditory features are grouped. Proximity is the basis for considering next-neighbor similarities and forming glimpses (for instance, local clusters of time-frequency units) under the assumption that elements which are located closely in time and frequency are likely to be dominated by the same object. As the auditory system seems to prioritize certain

cues in the case of inconsistent evidence, a crucial part of the grouping decision is to also estimate the reliability of each cue and define the cases in which some features may override others or should be given more weight.

The estimation of direct-neighbour similarities is supposed to facilitate the formation of local segments which can then be joined to form auditory streams. This tracking stage is yet to be implemented and requires a less local, but rather large-scale processing, taking more information into account than just that of the next neighbor. Grouping cues that come into play when taking a larger-scale perspective on the audio signals are common fate and closure. According to these principles, components that move at the same rate or into the same direction, whether it be location-wise, spectral, or in the form of synchronized onsets, can be grouped, and arrays of fragments that resemble a familiar template can be completed. Desirably, the implementation of these grouping principles will make it possible to blindly form clusters of similar elements and eventually assign them to physical sound sources.

Acknowledgements The authors would like to thank I. Koch and the two anonymous reviewers for their very helpful comments that have improved this chapter. This publication was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), SFB-1330, Hearing acoustics: Perceptual principles, Algorithms and Applications (HAPPAA).

References

- Alfás, F., J. Socorò, and X. Sevillano. 2016. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Science* 6 (5): 1–44.
- Anstis, S., and S. Saida. 1985. Adaption to auditory streaming of frequency-modulated tones. *Journal of Experimental Psychology: Human Perception and Performance* 11: 257–271.
- Arbogast, T., C. Maskon, and G. Kidd. 2002. The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America* 112 (5): 2086–2098.
- Assmann, P., and Q. Summerfield. 1994. The contribution of waveform interactions to the perception of concurrent vowels. *Journal of the Acoustical Society of America* 95: 471–484.
- Beauvois, M., and R. Meddis. 1991. A computer model of auditory stream segregation. *Quarterly Journal of Experimental Psychology* 43a: 517–541.
- Beauvois, M., and R. Meddis. 1996. Computer simulation of auditory stream segregation in alternating-tone sequences. *Journal of the Acoustical Society of America* 99 (4): 2270–2280.
- Bendixen, A., S. Denham, K. Gyimesi, and I. Winkler. 2010. Regular patterns stabilize auditory streams. *Journal of the Acoustical Society of America* 128: 3656–3666.
- Bey, C., and S. MacAdams. 2002. Schema-based processing in auditory scene analysis. *Perception and Psychophysics* 64: 844–854.
- Binder, M., N. Hirokawa, and U. Windhorst. 2009. *Encyclopedia of Neuroscience*. Berlin, Heidelberg: Springer.
- Bregman, A. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: The MIT Press.
- Brown, G., and M. Cooke. 1994. Computational auditory scene analysis. *Computer Speech and Language* 8: 297–336.
- Brungart, D., P. Chang, B. Simpson, and D. Wang. 2006. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of the Acoustical Society of America* 120 (6): 4007–4018.

- Brungart, D., P. Chang, B. Simpson, and D. Wang. 2009. Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers. *Journal of the Acoustical Society of America* 125 (6): 4006–4022.
- Cao, A., L. Li, and X. Wu. 2011. Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise. *Journal of the Acoustical Society of America* 129: 2227–2236.
- Carlyon, R., R. Cusack, J. Foxton, and I. Robertson. 2001. Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance* 27: 115–127.
- Carlyon, R., C. Plack, C. Fantini, and R. Cusack. 2003. Cross-modal and non-sensory influences on auditory streaming. *Perception* 32: 1393–1402.
- Chen, Z., and V. Hohmann. 2015. Online monaural speech enhancement based on periodicity analysis and a priori SNR estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23: 1904–1916.
- Cooke, M. 2006. A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America* 119: 1562–1573.
- Cooke, M., and F. Ellis. 2001. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication* 35: 141–177.
- Cusack, R., and B. Roberts. 2000. Effects of differences in timbre on sequential grouping. *Perception and Psychophysics* 62: 1112–1120.
- David, M., M. Lavandier, N. Grimault, and A. Oxenham. 2017. Discrimination and streaming of speech sounds based on differences in interaural and spectral cues. *Journal of the Acoustical Society of America* 142 (3): 1674–1685.
- Devergie, A., N. Grimault, B. Tillmann, and F. Berthomier. 2010. Effect of rhythmic attention on the segregation of interleaved melodies. *Journal of the Acoustical Society of America Express Letters* 128: 1–7.
- Dowling, W. 1973. The perception of interleaved melodies. *Cognitive Psychology* 5: 322–337.
- Drullmann, R. 1995. Speech intelligibility in noise: Relative contribution of speech elements above and below the noise level. *Journal of the Acoustical Society of America* 98: 1796–1798.
- Durlach, N., C. Mason, G. Kidd, T. Arbogast, H. Colburn, and B. Shinn-Cunningham. 2003. Note on informational masking(1). *Journal of the Acoustical Society of America* 113 (6): 2984–2987.
- Feldman, J. 2009. Bayes and the simplicity principle in perception. *Psychological Review* 116: 875–887.
- Froyen, V., J. Feldman, and M. Singh. 2015. Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological Review* 122: 575–597.
- Füllgrabe, C., and B. Moore. 2012. Objective and subjective measures of pure-tone stream segregation based on interaural time differences. *Hearing Research* 291: 24–33.
- Glasberg, B., and B. Moore. 1990. Derivation of auditory filter shapes from notched-noise data. *Hearing Research* 47: 103–138.
- Grimault, N., C. Micheyl, R. Carlyon, P. Arthaud, and L. Collet. 2000. Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency. *Journal of the Acoustical Society of America* 108: 263–271.
- Hartmann, W., and E. Fosler-Lussier. 2011. Investigations into the incorporation of the ideal binary mask in ASR. In *Proceeding of ICASSP*, 4804–4807.
- Hartmann, W., and D. Johnson. 1991. Stream segregation and peripheral channeling. *Music Perception* 9: 155–184.
- Hartmann, W., A. Narayanan, E. Fosler-Lussier, and D. Wang. 2013. A direct masking approach to robust ASR. *IEEE Transactions on Audio, Speech, and Language Processing* 21 (10): 1993–2005.
- Hu, G., and D. Wang. 2006. An auditory scene analysis approach to monaural speech segregation. *Topics in Acoustic Echo and Noise Control*, 485–515. Berlin, Heidelberg: Springer.
- Hu, G., and D. Wang. 2007. Auditory segmentation based on onset and offset analysis. *IEEE Transactions on Audio, Speech, and Language Processing* 15 (2): 396–405.
- Hu, G., and D. Wang. 2010. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing* 18 (8): 2067–2079.

- Jäkel, F., M. Singh, F. Wichmann, and M. Herzig. 2016. An overview of quantitative approaches in Gestalt perception. *Vision Research* 126: 3–8.
- Josupeit, A., N. Kopčo, and V. Hohmann. 2016. Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features. *Journal of the Acoustical Society of America* 139 (5): 2911–2923.
- Kidd, G., C. Mason, and V. Best. 2014. The role of syntax in maintaining the integrity of streams of speech. *Journal of the Acoustical Society of America* 135: 766–777.
- Li, N., and P. Loizou. 2008. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *Journal of the Acoustical Society of America* 123: 1673–1682.
- Li, Y., and D. Wang. 2008. On the optimality of ideal binary time-frequency masks. *Speech Communication* 51: 230–239.
- Lippmann, R. 1996. Accurate consonant perception without mid-frequency speech energy. *IEEE Transactions on Speech and Audio Processing* 4: 66–69.
- Mandel, M., R. Weiss, and D. Ellis. 2010. Model-based expectation maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing* 18 (2): 382–394.
- May, T., S. van de Par, and A. Kohlrausch. 2011. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech, and Language Processing* 19 (1): 1–13.
- McCabe, S., and M. Denham. 1997. A model of auditory streaming. *Journal of the Acoustical Society of America* 101: 1611–1621.
- Meddis, R. 1988. Simulation of auditory-neural transduction: Further studies. *Journal of the Acoustical Society of America* 83: 1056–1063.
- Moore, B. 2012. *An Introduction to the Psychology of Hearing*. Bingley, UK: Emerald Group. <https://books.google.de/books?id=LM9U8e28pLMC> (last accessed December 15, 2019).
- Moore, B., and H. Gockel. 2012. Properties of auditory stream formation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367: 919–931.
- Narayanan, A., and D. Wang. 2013a. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proceeding of ICASSP*, 7092–7096.
- Narayanan, A., and D. Wang. 2013b. The role of binary mask patterns in automatic speech recognition in background noise. *Journal of the Acoustical Society of America* 133 (5): 3083–3093.
- Ozerov, A., E. Vincent, and F. Bimbot. 2011. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*.
- Raj, B., M. Seltzer, and R. Stern. 2004. Reconstruction of missing features for robust speech recognition. *Speech Communication* 43: 275–296.
- Remez, R., P. Rubin, D. Pisoni, and T. Carrell. 1981. Speech perception without traditional speech cues. *Science* 212: 947–950.
- Roman, N., and D. Wang. 2006. Pitch-based monaural segregation of reverberant speech. *Journal of the Acoustical Society of America* 120 (1): 458–469.
- Roman, N., D. Wang, and G. Brown. 2003. Speech segregation based on sound localization. *Journal of the Acoustical Society of America* 114 (4): 2236–2252.
- Roman, N., and J. Woodruff. 2011. Intelligibility of reverberant noisy speech with ideal binary masking. *Journal of the Acoustical Society of America* 130 (4): 2153–2161.
- Schädler, M., A. Warzybok, S. Ewert, and B. Kollmeier. 2016. A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception. *Journal of the Acoustical Society of America* 139 (5): 2708–2722.
- Schoenmaker, E., and S. van de Par. 2016. Intelligibility for binaural speech with discarded low-SNR speech components. *Advances in Experimental Medicine and Biology* 894: 73–81.
- Schwartz, A., J. McDermott, and B. Shinn-Cunningham. 2012. Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *Journal of the Acoustical Society of America* 132 (1): 357–368.
- Shannon, R., F. Zeng, V. Kamath, J. Wygonsik, and M. Ekelid. 1995. Speech recognition with primarily temporal cues. *Science* 270: 303–304.

- Shinn-Cunningham, B., V. Best, and A. Lee. 2017. Auditory object formation and selection, 7–40. Simpson, S., and M. Cooke. 2005. Consonant identification in n-talker babble is a nonmonotonic function of n. *Journal of the Acoustical Society of America* 118: 2775–2778.
- Smaragdis, P., and J.C. Brown. 2003. Non-negative matrix factorization for polyphonic music transcription. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, 177–180. <https://doi.org/10.1109/ASPAA.2003.1285860> (last accessed December 15, 2019).
- Snyder, J., O. Carter, E. Hannon, and C. Alain. 2009. Adaptation reveals multiple levels of representation in auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance* 35 (4): 1232–1244.
- Spille, C., S. Ewert, B. Kollmeier, and B. Meyer. 2017. Predicting speech intelligibility with deep neural networks. *Computer Speech and Language* 48: 51–66.
- Srinivasan, S., N. Roman, and D. Wang. 2006. Binary and ratio time- frequency masks for robust speech recognition. *Speech Communication* 48: 1486–1501.
- Stainsby, T., C. Füllgrabe, H. Flanagan, S. Waldman, and B. Moore. 2011. Sequential streaming due to manipulation of interaural time differences. *Journal of the Acoustical Society of America* 130: 904–914.
- Sutojo, S., S. van de Par, and J. Thiemann. 2017. A distance measure to combine monaural and binaural auditory cues for sound source segregation. In *Proceeding of DAGA-17*, Dtsch. Ges. Akustik (DAGA), Berlin.
- Thompson, S., R. Carlyon, and R. Cusack. 2011. An objective measurement of the build-up of auditory streaming and of its modulation by attention. *Journal of Experimental Psychology: Human Perception and Performance* 37: 1253–1262.
- van Noorden, L. 1975. Temporal coherence in the perception of tone sequences. Ph.D. thesis, Eindhoven University of Technology, Eindhoven, Netherlands.
- Vliegen, J., B. Moore, and A. Oxenham. 1999. The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *Journal of the Acoustical Society of America* 106: 938–945.
- Wang, D., and G. Brown. 2006. *Computational Auditory Scene Analysis: Principles Algorithms, and Applications*. Hoboken, New Jersey: Wiley-IEEE Press.
- Wang, D., and G. Hu. 2006. Unvoiced speech segregation. In *Proceeding of ICASSP-06*, 953–956.
- Wang, Z., and D. Wang. 2016. Robust speech recognition from ratio masks. In *Proceeding of ICASSP-16*, 5720–5724.
- Warren, R., C. Obusek, and J. Ackroff. 1972. Auditory induction: Perceptual synthesis of absent sounds. *Science* 176: 1149–1151.
- Warren, R., K. Riener, J. Bashford, and B. Brubaker. 1995. Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Perception and Psychophysics* 57 (2): 175–182.
- Wertheimer, M. 1923. Untersuchungen zur Lehre von der Gestalt: II. *Psychologische Forschung* 4: 301–350.
- Woodruff, J., and D. Wang. 2010. Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization. *Transactions on Audio, Speech, and Language Processing* 18(7): 1856–1866.
- Woodruff, J., and D. Wang. 2013. Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues. In *IEEE Transactions on Audio, Speech, and Language Processing* 21 (4): 806–815.

Selective Binaural Attention and Attention Switching



Janina Fels, Josefa Oberem and Iring Koch

Abstract This chapter examines the cognitive control mechanisms underlying auditory selective attention by considering the influence of variables that increase the complexity of the auditory scene concerning technical aspects such as dynamic binaural hearing, room acoustics, head movements, and interfering noise sources as well as those that influence the efficiency of cognitive processing. Classical research in auditory selective attention does not represent realistic or close to real-life listening experiences, of which room acoustics, distracting sources, as well as the dynamic reproduction of an acoustic scene including head movements, are essential parts. The chapter starts with an introduction to the subject of maintaining and switching attention from the standpoint of cognitive psychology. A paradigm suitable for the study of intentional switching of auditory selective attention is introduced through dichotic stimulus representation with different single number words (1–9, excluding 5) uttered by speakers of different gender presented simultaneously, one to the participant's left ear and the other to the right ear. The listener is required to categorize, as quickly as possible, the target number as being either smaller or larger than five, with a visual cue indicating the listener's task in each trial. This paradigm is gradually extended from dichotic reproduction to a complex dynamic acoustic scene to study the binaural effects in selective attention and attention switching, including different room acoustic conditions. Various technical possibilities are evaluated to validate the binaural reproduction of the spatial scene, minimizing errors on account of the acoustic virtual reality. Additionally, the influence of different binaural reproduction methods on the selective attention and attention switching model is carefully examined and compared to a natural listening condition using loudspeakers in an anechoic setting. The application of the binaural listening paradigm in anechoic conditions tests a listener's ability to switch auditory attention in various setups intentionally. According to the results, intentional switching of the attention focus is associated

J. Fels (✉) · J. Oberem

Teaching and Research Area of Medical Acoustics, Institute of Technical Acoustics,
RWTH Aachen University, Aachen, Germany
e-mail: jfe@akustik.rwth-aachen.de

I. Koch

Institute of Psychology, RWTH Aachen University, Aachen, Germany

© Springer Nature Switzerland AG 2020

J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_3

with higher reaction times compared to maintaining the focus of attention on a single source. Also, particularly concerning the error rates, there is an observable effect of the stimulus category (i.e., stimuli spoken by target and distractor may evoke the same answer (congruent) or different answers (incongruent)). The congruency effect may be construed as an implicit performance measure of the degree to which task-irrelevant information is filtered out. The binaural paradigm has also been applied to older (slightly hearing-impaired) participants, with the results of which have been compared to experiments involving young normal-hearing participants, resulting in higher error rates and reaction times. Scenarios involving even more complex environments, including room acoustics (i.e., reverberation), have shown reaction times and error rates that rely significantly on reverberation time. Switch costs, in particular, reaction time differences between switch trials and repetition trials, can highly depend on the reverberation time.

1 Introduction and the Psychological Background

Many communicative situations involve multiple potentially competing sources of acoustic information that are simultaneously available for auditory processing. At a dinner party, for instance, a person can deliberately ignore the ambient noise and other conversations in the background to be able to listen to a friend's interesting story. Someones ability to stay focused on what the friend is narrating will depend on the capacity for attentional selection of the relevant acoustic information—see Shinn-Cunningham (2008) for an in-depth review.

In fact, listening to an individual speaker in a busy, boisterous setting is one of the best real-life examples of selective attention (Pashler 1998), which can be defined behaviorally as context-sensitive preferential stimulus selection in the presence of competing stimuli (i.e., voices). The cognitive mechanisms underlying selective listening entail a “bias” in auditory stimulus perception—either an “attentional set,” or a “task set” as described in Logan and Gordon (2001)—so that the relevant source of information can be filtered out while ignoring the irrelevant information.

1.1 *How It All Began*

Investigating selective auditory attention has had a long tradition in experimental psychology. A classic experimental paradigm for examining auditory selective attention is based on dichotic¹ listening (Broadbent 1958; Cherry 1953), in which different

¹In general, “binaural” refers to a presentation relating to two ears. The stimuli can either be identical (diotic) or different (dichotic)—e.g. see Blauert and Braasch (2008). In this chapter, the terms “binaural” and “dichotic” are used slightly differently from the standard definition. Here, “binaural” only refers to the situation where stimuli are presented to both ears and also include

information is presented simultaneously via headphones. Here, the instructions typically specify information presented to one ear as being task-relevant. In a pioneering study, Cherry (1953) presented two separate continuous speech messages to both ears of the participants, who were required to listen selectively and repeat immediately (“shadow”) the speech presented to one ear while ignoring the task-irrelevant speech presented to the other ear. The speech on the irrelevant side (i.e., ear) always began with an English utterance spoken by a male voice and ended with an English utterance, while the middle portion of the speech differed across experimental conditions. Following the shadowing task, participants were asked whether they could recall the contents of the irrelevant information, or whether they had noticed anything unusual about it. It was found that, while the subjects could identify the irrelevant information as speech, their memory of the content was surprisingly poor. They often failed to notice any changes in the middle, such as a switch from English to German, or a switch to a backward-played speech condition. The changes in the gender of the task-irrelevant speaker, however, had been largely noticed by the participants. Since then, numerous studies have proved the existence of auditory selection capabilities in healthy individuals, with task-relevant information processed successfully and task-irrelevant information suppressed—e.g., see the reviews of Bronkhorst (2015), Hugdahl (2011), and Pashler (1998).

Regarding the nature of this processing selectivity, a prominent theoretical account on “early selection” or “filter theory” (Broadbent 1958) postulates that an attentional filter operates on the perceptual level prior to semantic processing. Thus, while perceptual “surface” features of the to-be-ignored speech, such as the gender of the speaker, can be encoded in two concurrently available auditory streams, the attentional filter enforces strictly serial processing of semantic information. Other views, however, consider parallel processing of competing information all the way up to post-perceptual, semantic processing levels (e.g., Treisman 1969)—for a review, see Pashler (1998). This question has been extensively examined owing to its theoretical relevance to the cognitive processes underlying attentional selection, and thus, dichotic listening has become a preferred experimental paradigm.

1.2 *Maintaining and Switching Attention*

As described earlier, in a dichotic-listening experiment, participants are required to attend to information presented to one ear while ignoring information presented to the other ear. The question of how much nominally unattended information is actually processed is important given the postulation of the early filter theory of attention (e.g., Broadbent 1958)—for a review, see Holender (1986). According to this theory, attentional selection occurs prior to any semantic processing of the

spatial information. The term “dichotic” only refers to two different stimuli presented monaurally to the opposite ears.

irrelevant information. Thus, any evidence of semantic processing of unattended information would potentially refute the filter theory.

However, empirical evidence of simultaneous semantic processing of both sound sources has been repeatedly reported—see Pashler (1998) for a historically oriented review. One way to reconcile the notion of early selection with the evidence of seemingly parallel processing of semantic information is to assume that rapid serial attention switching can occur between sound sources. For instance, when instructed to respond to the stimulus presented to the left ear, participants may inadvertently process the stimulus at the irrelevant location by switching attention swiftly back and forth between the two ears—see Lachter et al. (2004) and Lavie (2005) for discussions.

Thus, the supposedly semantic processing of truly unattended information may be due to attention switches. Substantial research effort has been invested to address this methodological issue in the prevention of involuntary attention switches (Wood and Cowan 1995a, b). Rivenez et al. (2008) provide a taxonomy of studies on this topic. However, this view of the “structural inability” of human auditory information processing to encode semantic information from disparate sound sources in parallel has neglected the cognitive processes that give human information processing its substantial flexibility.

The focus on involuntary switches (i.e., “capture”) of auditory attention has resulted in a paucity of knowledge with respect to the mechanisms that actually enable the listener to flexibly switch attention from one stream of information to another, as in typical multi-speaker situations—for reviews, see Bronkhorst (2015) and Shinn-Cunningham (2008). To examine the processes underlying intentional switching of attention in selective listening situations, a novel version of the dichotic listening paradigm has been developed by Koch et al. (2011) based on the cuing version of task switching—for reviews, see Monsell (2003), Koch et al. (2018), Kiesel et al. (2010), and Vandierendonck et al. (2010).

Unlike many earlier studies of auditory selective attention, which measured speech perception and comprehension directly in terms of the accuracy of the report, the model described here assesses auditory attention by requiring participants to categorize a target as quickly as possible while ignoring a simultaneously presented distractor stimulus. This implicit speeded “online” reaction-time task helps measure performance both in terms of reaction time and accuracy. This basic experimental paradigm and recent behavioral findings are described in the next section, followed by a discussion of the technical development of this paradigm to extend it to binaural listening situations which may be auralized in virtual settings and validated using performance measures developed in the basic dichotic setup of this study.

2 The Dichotic Paradigm to Study Intentional Switching of Auditory Attention

The basic paradigm employs the so-called task-cuing method. The specific task and its associated cognitive processing operations can vary from trial to trial, with the required task being indicated prior to each trial by an instruction cue as described by Meiran (1996). Jost et al. (2013) provide a review of cue processing in task switching. In our cued auditory attention-switching paradigm (Koch et al. 2011), two different number words are presented simultaneously, one to the participants’ left ear and the other to the right ear (i.e., dichotic listening), with the target number requiring to be identified as quickly as possible as being either smaller or larger than five, using a left and right response key, respectively. The cue indicates the listener’s task in each trial (see Fig. 1). It must be noted, however, that the term “task” is not easy to define given that tasks can be defined at different levels. Here, the attentional selection criterion (i.e., left or right ear) is designated as a task, and therefore a switch to the relevant ear would correspond to what is termed a “task switch” in the literature on task switching—for reviews, see Kiesel et al. (2010) and Koch et al. (2018). In the authors’ cuing paradigm, the ear, where the relevant speaker is presented, is indicated prior to each trial by an explicit visual task cue. Thus, it may be described as a task-switching paradigm, which helps isolate one particular task processing component, namely the biasing of attention for (auditory) target selection—see, e.g., Logan (2005), and Logan and Gordon (2001). It has allowed examining essential aspects of auditory attention in selective listening to a single voice in the presence of

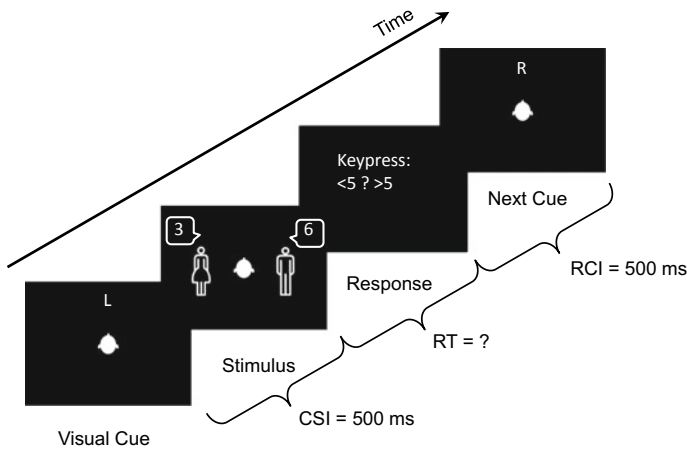


Fig. 1 Trial procedure in a dichotic setup with a visual cue indicating the target direction, a Cue-Stimulus Interval (CSI) of 500 ms, the synchronous presentation of the stimuli, reaction time between onset of stimulus and the response of the participant, and the Response-Cue Interval (RCI) of 500 ms

multiple speakers, as in the “cocktail party” situation (Bronkhorst 2015; Koch et al. 2011), while at the same time affording close experimental control.

While previous studies have indicated selective listening benefits of cuing with respect to the spatial target position in complex multi-talker situations (Brungart and Simpson 2004; Kidd et al. 2005a; Kitterick et al. 2010), there are significant methodological differences between those studies and the cued auditory attention-switching paradigm used here as described by Koch et al. (2011), and Lawo and Koch (2014a). Most notably, previous studies used measures of perceptual accuracy instead of reaction time measures, and, also, did not explicitly focus on the issue of flexibly switching attentional settings.

2.1 Switch Costs, Congruency Effect, and Temporal Dynamics

It was examined whether instructed, intended changes in the auditory selection (or “filter”) criterion would incur performance costs using the cued auditory attention-switching paradigm. The performance was found to be significantly worse in a task switch than during a repetition, in which the target-defining feature remained the same (Koch et al. 2011). This switch cost points to cognitive interference in information processing when the selection criterion needs to be intentionally adjusted (see Fig. 2).

The current paradigm also helps examine whether the irrelevant auditory information is nevertheless encoded, creating interference in the processing of task-relevant information. This interference is measured as impaired performance when the numer-

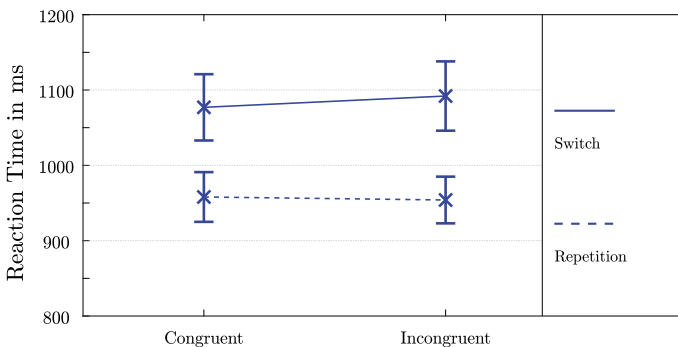


Fig. 2 Reaction time (in ms) as a function of attention switch and congruency. Error bars indicate standard errors. The main effect of attention switch is significant and amounts to switch costs of 126 ms. The main effect of congruency and the interaction was not significant. In error rates (not shown in the graph), the congruency effect turned out to be highly significant. In further studies, the congruency effect was also significant in reaction times. Note that in this experiment, the selection cue was the speaker’s gender and not the ear of presentation. For more detailed information, see Koch et al. (2011)

ical category of the relevant auditory information (>5 or <5) is different (i.e., “incongruent”; e.g., presenting acoustically the number “*three*” to the left ear and the number “*six*” to the right ear) compared to when the categorization is the same (“congruent”; e.g., playing the number “*three*” to the left ear and the number “*four*” to the right ear) for both auditory inputs. This congruence effect (e.g., Kiesel et al. 2010) may be construed as an “implicit” performance measure of attending to task-irrelevant information (i.e., disobeying task instructions), and the switch costs as an explicit measure of how well instructions were followed to switch attention. In fact, in previous studies, it was found the performance to be worse when the non-target number was incongruent with the target number, thus requiring a different behavioral key-press response relative to congruent numbers. This congruence effect represents the influence of the processing of irrelevant information, whereas the attentional switch costs supposedly index the cognitive control processes involving the reconfiguring (“biasing”) of attention to process relevant information based on a new selection criterion—see Koch et al. (2010), Koch and Lawo (2014), and Lawo et al. (2014).

In another study, Lawo et al. (2014) examined the role of the stimulus dimension on which selection is required. Their participants selected the target either based on the speaker’s gender (as in Koch et al. 2011) or by the ear of presentation. The distinction between gender-based and ear-based auditory target selection could be related to the difference between processing along the frequency dimension and the spatial dimension. Selection by spatial location plays a primary role in visual attention, but the neurophysiology of audition, which is characterized by tonotopic coding (instead of retinotopic coding in visual processing), suggests that frequency coding is likely more preponderant than spatial processing (e.g., see Woods et al. 2001; Shinn-Cunningham 2008). Hence, it is important to examine whether spatial selection (by ear) would be beneficial relative to selection according to the speaker’s gender. In fact, Lawo and Koch (2014a) found that switch costs were even greater for ear-based selection than for gender-based selection. Notably, however, on repetition trials, the performance was actually better with ear-based selection (particularly with long cuing intervals, whereas there was no such benefit in switch trials, see below). This observation suggests that ear-based selection can be highly efficient, although situations requiring flexible attention shifting diminish this relative efficiency.

In addition to these basic findings with respect to the switch costs, the role of the type of selection criterion (speaker’s gender vs. speaker’s location), and the processing of irrelevant information (i.e., the congruency effect), we have also examined the temporal dynamics of the cognitive control processes in such intentional auditory attention switches. The use of a specific, explicit instruction cue made it possible to examine the influence of the attentional processing time course. There are two relevant temporal intervals in this paradigm.

First, there is the interval between the response in the last trial and the presentation of the next task cue—the so-called Response Cue Interval, RCI. Because the identity of the upcoming cue is not predictable, participants could not know whether the next speaker to attend to would be male or female (or left or right ear, respectively). The general observation, with respect to the switch costs in this situation, suggests that participants stay “tuned” to the previously relevant speaker (or the speaker category,

because different female and male voices are used). Thereby, they incur a performance cost when this processing bias needs to be re-adjusted (i.e., a new selection criterion to be implemented in the cognitive processing of the two competing auditory signals). The RCI represents the period during which this bias gradually “dissipates”. However, RCI variations do not affect performance, at least within relatively small time frames (up to a 1,000 ms), suggesting that auditory attention settings persist to a certain extent, facilitating the processing of auditory information that falls into the previously relevant category (Koch and Lawo 2014). However, whether this facilitation would persist over somewhat longer time ranges of several seconds remains to be tested.

Second, the interval between the cue and the subsequent auditory target stimulus (Cue-Stimulus Interval, CSI) can be used to prepare for a switch in the selection criterion. According to the authors’ observation in several studies, CSI variations have little influence on performance costs due to attention switching (Koch et al. 2011; Lawo and Koch 2015; Lawo et al. 2014), indicating the difficulty of preparing for a new auditory target selection criterion prior to the onset of the auditory signal itself. More recently, however, it was found that such a preparation can be more successful if the attention switch is not indicated by a cue in an otherwise unpredictable sequence of attention switches but, instead, by a pre-instructed, memorized sequence. This suggests that, instead of “exogenous” cues, endogenous prediction of the relevant criterion for selecting the next auditory target may be more pertinent for auditory attentional preparation (Seibold et al. 2018).

Taken together, the results of these studies using the task-switching version of dichotic listening show several essential features of auditory attention in multi-talker situations. First, although participants can easily follow the instruction to switch auditory attention to a new target, this switching results in performance costs in terms of increased reaction times and reduced response accuracy. Second, while the participants mostly succeed in responding to a new auditory target and thus listen selectively to the relevant information, they cannot avoid processing the irrelevant information to the extent that the irrelevant distractor stimulus does not influence their response (i.e., the congruency effects). Third, selecting the auditory target in situations that require spatial target selection (i.e., left vs. right ear) does not decrease the performance costs relative to attention switches based on the gender of the relevant speaker. This occurs despite the greater benefit of attending to the same location repeatedly compared to repeatedly attending to the same speaker gender (so that the switch costs are even higher with spatial selection). Fourth, auditory attention represents a temporally stable cognitive setting that does not either passively dissipate quickly (within a second or so) or easily changes in preparation for an auditory target switch, indicating some degree of “auditory attentional inertia”.

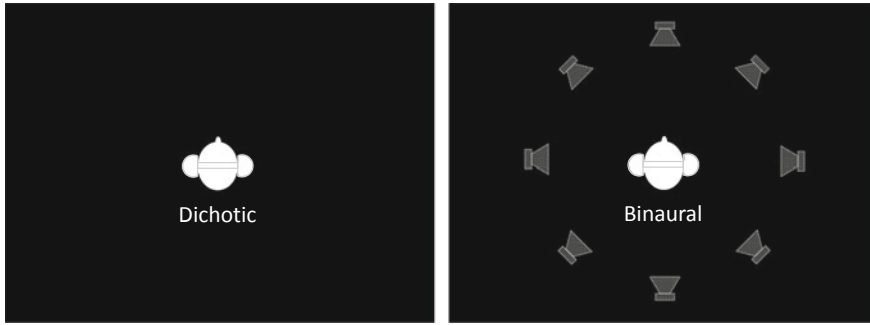


Fig. 3 Sketch of dichotic and binaural reproductions. In a dichotic reproduction, the represented stimuli are not perceived spatially, but are located in or close to the head on the right and left sides. In a binaural reproduction using head-related transfer functions (HRTFs), the presented stimuli are mostly externalized and located in space. Grayed loudspeaker symbols represent possible locations in space

2.2 Constraint of Dichotic Listening

In addition to being technically easy to handle, the described auditory attention-switching version of the dichotic listening paradigm uses experimentally well-controlled stimuli and is capable of exact performance measurement (with a high resolution at the level of reaction time in milliseconds as well as error rates). However, despite these attractive methodological features, which help to isolate the influence of experimental variables, it is clear that dichotic listening represents a rather unnatural situation, whereas ordinary selective listening situations (e.g., a conversation in a restaurant) also include a number of additional cues that are associated with binaural hearing—see Fig. 3.

3 Exploring Auditory Selective Attention and Attention Switching Through Binaural Reproduction

3.1 The Binaural Paradigm

Based on observations for the use of dichotic stimulus presentations, the authors have recently developed a binaural paradigm to study attention switching in selective listening situations (Oberem et al. 2017, 2014, 2018). In this context, the term *binaural* refers not only to the situation in which sound reaches both ears but also to spatial information.

The basic binaural paradigm consists of two simultaneously presented stimuli, which are delivered by two speakers of the opposite gender. The speakers are located

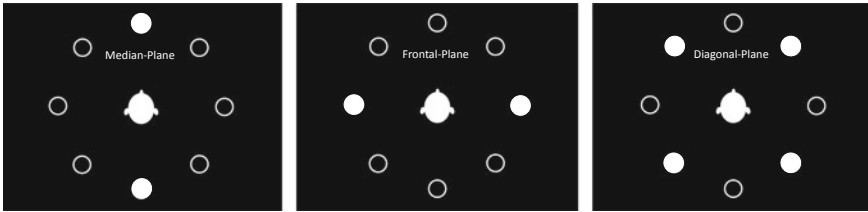


Fig. 4 The paradigm offers eight possible locations on the horizontal plane for target and distracting speakers. To analyze the influence of these positions, they are arranged in three groups. The positions in front and back are on the median plane, the ones to the left and right are on the frontal plane, and the remaining four are on a diagonal plane

at two different positions, out of eight possible locations around the listener. These positions are evenly distributed on the horizontal plane (see Fig. 4).

In each trial, the number word spoken by one speaker is the target, while the other one represents the distractor. The participant is asked to focus on the target speaker and ignore the distracting speaker. The target speaker’s location is indicated in advance through a visual cue shown on a monitor to distinguish between target and distractor. The visual cue consists of a sketch of all directions indicating the target direction with a filled dot. The listener’s task is to categorize the stimulus of the target speaker (i.e., the spoken number word) into less than vs. greater than five. The two stimulus categories are mapped to two response buttons, held in hand, to be pressed by the left or the right thumb.

Each trial starts with a visual cue presented on the monitor in front of the participant. After a cue-stimulus interval of 500 ms, the two acoustic stimuli (target and distractor) are presented simultaneously, with the visual cue remaining on the screen until the participant’s response to the acoustic target. The interval between the response and the next cue is also set to 500 ms. In case of an error, visual feedback (“Fehler”, German for “error”) is displayed for 500 ms, delaying the onset of the next cue (see Fig. 5).

The stimuli, the number of distractors, the task, and the equipment are adjusted according to the focus and requirements of the study (Oberem et al. 2017, 2014, 2018).

3.2 *Authenticity and Plausibility in Binaural Reproduction*

To simulate a real-life condition, the (binaural) reproduction of the spatial scene needs to be adequately plausible, if not authentic. Blauert (1997) defines the perceptual identity in a comparison between a real scene and a virtual scene as “authentic”. A “plausible” reproduction, on the other hand, refers to a scenario in which the perceptual identity is not essential, which, according to Lindau and Weinzierl (2012)

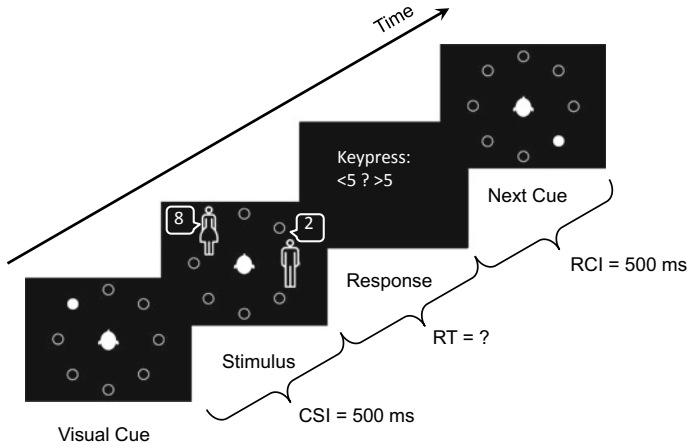


Fig. 5 Procedure of a trial in a binaural setup with a visual cue indicating the target direction, a Cue-Stimulus Interval (CSI) of 500 ms, the synchronous presentation of the stimuli, reaction time between onset of stimulus and the response of the participant, and the Response-Cue Interval (RCI) of 500 ms

can be understood by “a simulation in agreement with the listener’s expectation towards a corresponding real event”.

Thus, it is crucially important to examine how spatial reproduction affects the results in experiments involving auditory selective attention to determine whether different conclusions may be reached with different spatial reproduction methods or not. In this context, it may be pertinent to ask what qualitative difference, in terms of results, is likely between “real-life” situations simulated with high accuracy (including individual head-related transfer functions, etc.) and merely plausible rendering methods.

Measurement paradigms for auditory selective attention may also serve as indices of binaural reproduction quality, especially concerning the quality of acoustic virtual reality situations such as a cocktail party.

In Oberem et al. (2016), different methods of reproducing binaural stimuli were examined in terms of their authenticity and plausibility. Two different microphone setups were tested. The miniature microphones were either placed in a small silicon fixture to create an open dome or they were inserted in earplugs to block the ear canals. These setups were then compared in a real-source scenario using individual head-related transfer function (HRTF) and headphone transfer function (HpTF) measurements. They were tested in an anechoic chamber using loudspeakers with the quality of the binaural reproduction via headphones (compare Fig. 6). Using a robust equalization paradigm, as in Masiero and Fels (2011) where headphones were repositioned by the participants after each of the eight HpTF measurements, equalization curves were calculated using the mean of the HpTF measurements.

Two listening experiments involving 80 participants were conducted with a focus on authenticity and plausibility (Oberem et al. 2016). In an indirect comparison

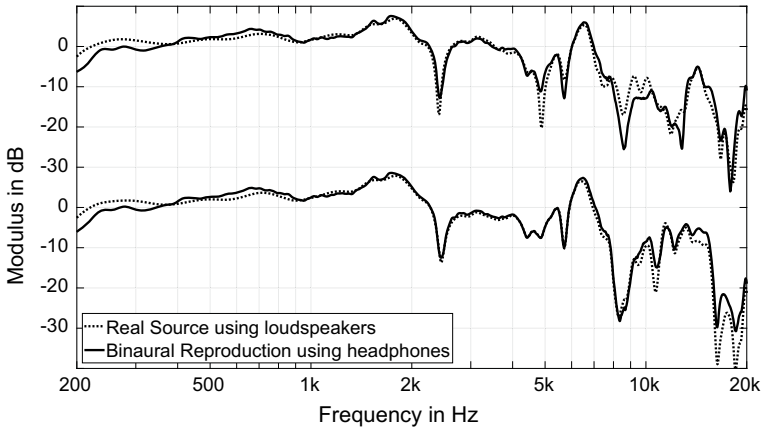


Fig. 6 Real and virtual HRTF measured with a blocked meatus (upper graph) and an open meatus (lower graph). The measurements of “real” HRTFs conformed to the usual approach of HRTF measurements. The binaurally synthesized stimulus was presented via headphones and recorded with the microphone positioned at the entrance of the ear canal to measure “virtual” HRTFs. The recordings were divided by the original excitation signal to obtain a transfer function. For perfect binaural reproductions, the recorded signals were supposed to be identical. For more detailed information, see Oberem et al. (2016)

(plausibility), participants were not able to identify the reproduction system (real loudspeakers vs. binaural synthesis), although the stimulus was pulsed pink noise. In a direct comparison concerning authenticity, the performance was found to be highly dependent on the stimulus (speech, music, and pulsed pink noise). The coloration could often distinguish pink noise in higher frequencies and relatively small differences in location. In this study, no significant difference was observed between the HRTF/HpTF measurements with open dome and ear plug.

As the results of this investigation demonstrated, individual binaural reproduction with state-of-the-art methods in HRTF and HpTF measurements are largely plausible and, therefore, can be used in psycho-acoustic experiments or experiments seeking to assess psychological effects like auditory attention in which HRTF and HpTF measurements and the listening test are conducted separately.

The findings of this study are in line with those by Hartmann and Wittenberg (1996), who reported that their participants were not able to differentiate between real sources and the binaural reproduction when a synthesized vowel was used as a stimulus. Furthermore, Zahorik et al. (1996) described that listeners were unable to discriminate between reproduction sources and noise bursts. Langendijk and Bronkhorst (2000) has also conducted studies in plausibility, Moore et al. (2010), and Schärer and Lindau (2009).

3.3 Validating Binaural Quality with Well-Established Distance and Localization Tasks

It is possible to create and study complex acoustic scenarios using binaural reproductions, including the physically correct ear-canal input signals. The use of acoustic virtual reality can easily manipulate the location of sources in the room, the distance to the listener, and the influence of room acoustics and maskers. Distance and localization tasks can help validate the quality of a binaural reproduction.

Studies regarding the distance between sources as well as the distance between sources and the listener have been conducted by Best et al. (2005, 2007, 2010), Kidd et al. (2005b), Allen et al. (2009), and Mondor et al. (1998). In most of these studies, non-individual head-related transfer functions (HRTFs) obtained with the help of artificial heads were used to create the stimuli using binaural synthesis. It is often overlooked, however, whether the results of an experiment using real sources are significantly different from those of one that employs virtual sources.

Localization performance comparisons between real sources and individual binaural syntheses presented with headphones were analyzed and rated as similar by Bronkhorst (1995). While Wightman and Kistler (1989) found similar results, they also reported challenges for the individual binaural synthesis in elevated positions, which became apparent through an increased angle error. Several authors, namely Searle et al. (1975), Butler and Belendiuk (1977), Wenzel et al. (1993), and Møller et al. (1996), all of whom found that individual recordings had yielded better results compared to non-individual recordings. Detailed results also revealed that, in localization tasks, non-individual binaural stimuli caused difficulties for sources located in the median plane, on cones of confusion, as well as elevated directions.

In real-life scenes, participants are usually required to process much more complex information than in simple localization tasks. Hence, the aim was to find a new measure to define the required accuracy of binaural syntheses in an *everyday task* that would include localization, but the focus would be on a non-localizing component. The application of the paradigm on intentional switching in auditory selective attention (Koch et al. 2011) describes a search and categorization task. To successfully comply with the task, the participant needs to localize the target's speakers position correctly. Whether the task is performed correctly by the participant becomes observable in error rates. Reaction times give further information about the complexity of the task and therefore, indirectly about the localization performance.

3.4 The Role of Binaural Reproduction Methods in a Paradigm for Investigating Intentional Switching in Auditory Selective Attention

As described in Sect. 3.1, and in greater detail in Oberem et al. (2014), the previous paradigm is extended to a more natural and realistic setup by changing it to a

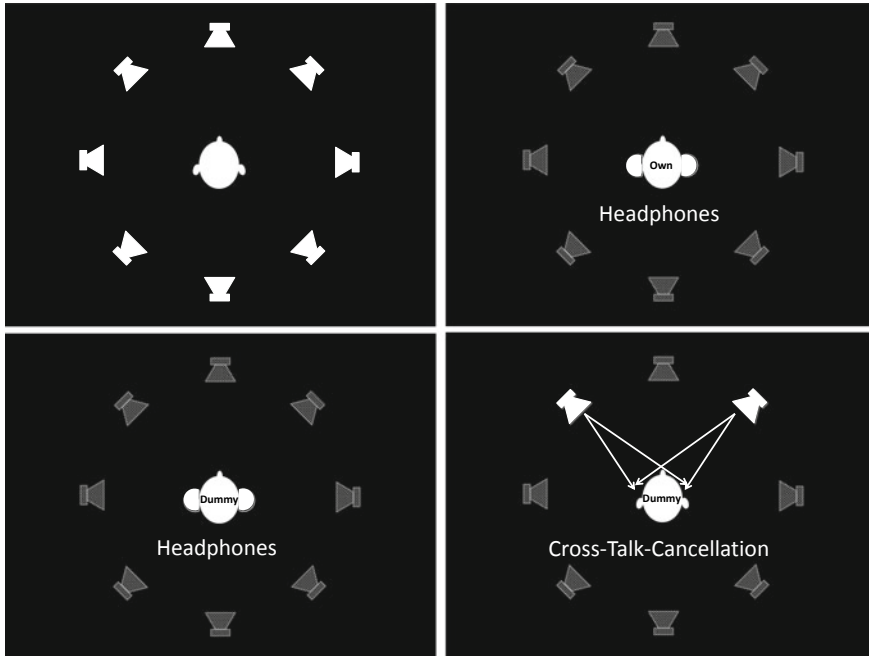


Fig. 7 Sketch of four different binaural reproduction methods to obtain a spatial scene. Several real sources (e.g., using loudspeakers in an anechoic chamber) can represent a spatial scene. A spatial scene can also be reproduced via headphones. Therefore, individual HRTFs or HRTFs from a binaural manikin need to be measured and convoluted with the stimulus to get a spatial impression. Using CTC filters, a binaural scene can be reproduced with at least two loudspeakers. For more detailed information, see Oberem et al. (2014)

binaural listening paradigm. The same experiment was repeated with four different reproduction methods to obtain a spatial scene: real sources (i.e., loudspeakers) in an anechoic environment, individual binaural synthesis reproduced with headphones, non-individual binaural synthesis reproduced with headphones, and non-individual binaural synthesis reproduced with two loudspeakers using Crosstalk Cancellation Filters (CTC) as shown in Fig. 7.

Not only localization ability suffers from non-individual binaural reproduction, but also reaction times and error rates in tasks of auditory selective attention. The absolute values of reaction times and error rates obtained in this study of auditory selective attention increase, significantly to an extent, with the individuality of the reproduction method. Thus, the shortest reaction times and lowest error rates were found for the real-source condition and the longest reaction times and highest error rates for the CTC condition (Fig. 8).

As expected, the reaction times and error rates were found to be higher concerning conditions using non-individual binaural synthesis. As studies by Searle et al. (1975), Butler and Belendiuk (1977), and Møller et al. (1996) show, localization also

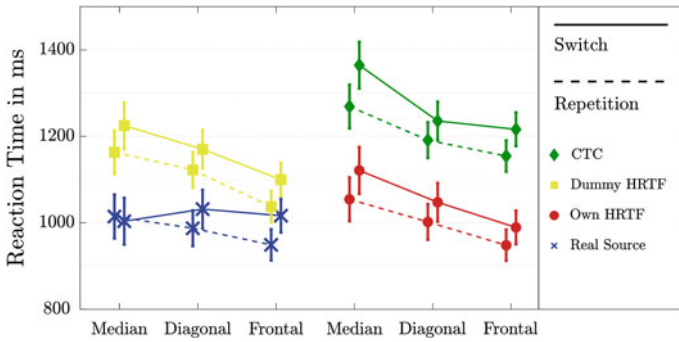


Fig. 8 Reaction time (in ms) as a function of reproduction method, target’s spatial position, and attention switch. Error bars indicate standard errors. The main effect of reproduction is significant, yielding significant differences between the individual methods (Real Source, crosses connected by blue lines, and individual HRTFs, dots connected by red line) and the non-individual methods using HRTFs measured with a binaural manikin (squares connected by yellow lines) and CTC (diamonds connected by green lines). The main effect of the target’s position is shown on the x-axis, describing the target’s speaker position in space on the median, diagonal, or frontal planes (Fig. 4). Post-hoc test shows significant differences between all three positions. The main effect of attention switch can be seen in solid and dotted lines. Independent of the reproduction method, participants react significantly faster when the target’s position is repeated and not switched. This graphic does not show all observed effects (e.g., congruency), for more information, see Oberem et al. (2014)

suffers from non-individual binaural stimuli compared to stimulus material based on individual HRTFs. Interestingly, the loss of individual information does not only hinder correct mapping of the source position in space, but it also impedes, as the error rates and reaction times prove, cognitive processing and attention tasks. The differences between non-individual binaural syntheses reproduced with headphones and those rendered via CTC occur due to higher degrees of freedom afforded by the latter. In CTC evaluations concerning localization (Gardner 1997; Takeuchi et al. 2001; Lentz et al. 2005; Bai and Lee 2006) limited sweet spots raised a challenge and affected performance negatively. The reproduction method with CTC resulted in the longest reaction times and the highest error rates in this study.

While the reaction times in the reproduction method with real loudspeakers did not differ significantly from those in the method with individual HRTFs, an unexpected but significant difference was found in the error rates. The difference between the reproduction conditions with real loudspeakers and individual HRTFs was the static presentation of the binaural synthesis. In both reproduction methods, participants were allowed to perform small head movements (Freedman and Fisher 1968; Iwaya et al. 2003; Jongkees and D. Veer 1958; Perrett and Noble 1997a,b; Thurlow et al. 1967; Wallach 1940; Young 1931; Toshima and Aoki 2006) within the area defined by the tracker. While the participants listening to the real sources benefited from the changes in interaural level difference (ILD) and interaural time difference (ITD) due to the small head movements, those listening to the static binaural synthesis missed this additional localization information. This lack of advantage concerning

the additional localization information might have been partly responsible for the latter group's increased error rates. For greater detail see Oberem et al. (2014).

3.5 Switch Costs and Congruency Effect in Binaural Experiment

Performance costs induced by attentional switches were observed in dichotic listening experiments by Lawo and Koch (2014a). Participants were also found to respond more slowly when the target's direction was switched in binaural listening, with the switching cost providing an explicit measure of how well instructions to switch attention could be followed. In general, the switch costs were found to be greater in dichotic listening experiments compared to the binaural investigation (~100 ms vs. 55 ms)—see Fig. 8. A switch between only two possible directions (i.e., dichotic listening) was expected to be easier to detect than a switch to one of eight possible directions equally distributed on the horizontal plane. The angular distance between the target's positions could have been a reason for different switch complexities. Besides the angular distance of the target's positions, the visual cue (in ear-based dichotic experiments, the visual cue was a letter (L/R) and therefore differed from the cue design of this investigation) might have had an effect on the switch costs—for more detail, see Oberem et al. (2014).

The effects of stimulus congruency showed the same patterns in binaural and dichotic listening tasks (Koch et al. 2011; Lawo and Koch 2014a) and could be construed as implicit performance measures of attending to task-relevant information and filtering out the irrelevant information (Koch et al. 2011). In the binaural paradigm, the congruency effect was found to be most distinct when real sources were used. Thus, the distracting information was less effectively ignored when the reproduction method was based on binaural synthesis relative to the real sources. The loss of the additional localization information due to a static reproduction could have been a reason for this effect.

A more complex binaural listening paradigm allows the analysis of additional effects such as the spatial combination of target and distractor's location. Target's and distractor's positions may be wide apart, directly neighbored or within one cone of confusion. Longer reaction times and higher error rates were found in the latter conditions (cf., Oberem et al. 2014), and these effects have also been observed in localization experiments. Using real sources, individual and non-individual reproductions via headphones, Møller et al. (1996) found errors to accumulate within the cone of confusion, especially the median plane. In the non-individual reproduction cases, in particular, the percentage of errors in the median plane conditions was seen to increase. The effect of spatial separation of sources in experiments focusing on selective attention was studied by Best et al. (2006), who found that auditory selective attention was worse when sources were not (or only slightly) spatially separated.

These observations, however, were based only on error rates as the paradigm by Best et al. (2006) did not allow the measurement of reaction times.

According to this investigation, the extension of a dichotic paradigm to a binaural one affords greater opportunities to analyze intentional switching in auditory attention. The comparison of reproduction methods showed that the differences between absolute values of reaction time and error rates should not be neglected. In experiments where the effects and interactions of different variables were compared, all reproduction methods were found to yield almost identical results.

4 Exploring Age-Related Effects on Intentional Switching of Auditory Selective Attention in a Spatial Setup

Using the binaural listening paradigm described in Sect. 3.1, age-related differences in the ability to intentionally switch auditory selective attention between two speakers defined by their spatial location were examined. The results of 20 young normal-hearing (\bar{x} 24.8 years) and 20 older normal-hearing to slightly hearing-impaired (\bar{x} 67.8 years) participants were compared. The spatial reproduction of stimuli was achieved by headphones using non-individual head-related transfer functions of an artificial head (Oberem et al. 2017).

A comparison between the two groups of participants revealed group differences in terms of absolute values of reaction times and error rates (Fig. 9). These (expected) results were in line with the previous dichotic investigation of age-related effects in intentionally switching auditory selective attention (Lawo and Koch 2014b). Increased reaction times and error rates with respect to older participants have also been found in other dichotic and binaural investigations involving attention (Abel et al. 2000; Dobрева et al. 2011; Duquesnoy 1983; Getzmann et al. 2015; Helfer et al. 2013; Humes et al. 2006; Kramer et al. 1999; Kray et al. 2008; Li et al. 2004; Marrone et al. 2008; Singh et al. 2013; Tun et al. 2002).

A significant effect was seen for the endogenous² attention switch, indicating that participants responded faster when the target's direction was repeated, was observed in both age groups in the present study as well as in previous investigations using dichotic and binaural listening (Koch et al. 2011; Lawo et al. 2014; Oberem et al. 2014). Switch costs, which provided an explicit measure of how well instructions to switch attention could be followed, did not differ significantly from those of the previous binaural investigation (Oberem et al. 2014). The inhibition of competing perceptual filter settings may be important for success in the attention switching task. That there is an age-related decline in the ability to inhibit irrelevant information has been predicted in several theories (Braver and Barch 2002; Hasher

²In contrast to exogenous cues, which are often used in detection tasks and lead to automatic (i.e., bottom up) target selection, we used endogenous cues (e.g., visual symbolic cue at screen center) that need attention to “actively” select (i.e., top down) the target stimulus before the categorization task could be performed.

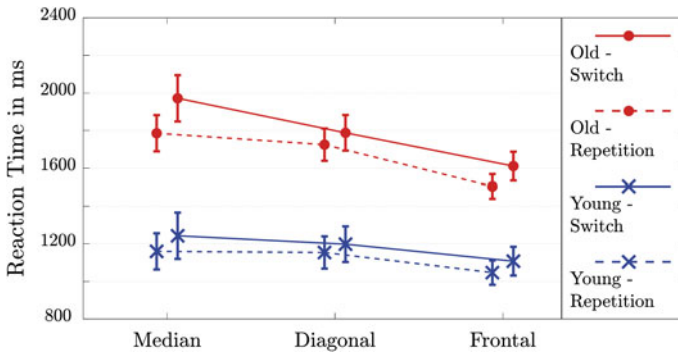


Fig. 9 Reaction time (in ms) as a function of age, target's spatial position, and attention switch. Error bars indicate standard errors. The main effect of age in reaction times between young (crosses connected by blue lines) and old (dots connected by red lines) participants amounts to 580 ms. The main effect of the target's position is shown on the x-axis, describing the target's speaker position in space on the median, diagonal or frontal plane (Fig. 4). The post-hoc test shows significant differences between all three positions. The main effect of attention switch can be seen in solid and dotted lines. Older and younger participants react significantly faster when the target's position is repeated and not switched. For more information, see Oberem et al. (2017)

et al. 2001). In this investigation, however, the auditory switch costs in older participants proved to be similar to those in younger participants, indicating no age-related differences in attention switching (Fig. 9). This observation corroborates the findings of previous research using a simpler dichotic listening set-up (Lawo and Koch 2014b). Assuming that inhibitory processes contribute to auditory switch costs, the results of this investigation deviated from the inhibitory deficit theory. In simple tasks involving exogenous attention switches, Singh et al. (2013) also did not observe any age-related deficits concerning word identification scores. However, significant age-related attentional deficits have been detected in more complicated tasks involving multiple attention switches. In terms of error rates, these findings agree with those of the present study. While the observation of non-significant age-related differences in switch costs in terms of reaction times (in addition to error rates) represents a null effect, it nevertheless provides additional evidence that the performance of older participants is similar to their younger counterparts in intentional attention switching tasks, despite the general age-related slowing of responses.

The ability of younger and older participants to focus their attention on one speaker and simultaneously ignore the distracting speaker was analyzed by examining the congruency of number words. The most significant difference between the dichotic investigation (Lawo and Koch 2014b) and the present binaural investigation was found in the interaction of congruency and the age-related effect. The present investigation showed a significant variation in reaction times, indicating that older adults performed comparatively worse when the stimuli were incongruent, which was not seen in the dichotic investigation. The difference between congruent and incongruent trials in reaction time was three times greater for older participants than for young

adults. It may, therefore, be assumed that, in binaural listening situations, older people have more difficulty ignoring a second speaker compared to their younger counterparts. Thus, the current results appear to be in line with the hypothesis that older adults suffer from a deficit in inhibitory processes (Braver and Barch 2002; Hasher et al. 2001). Considering that there was no age-related effect in attention switch costs, it may be assumed that the ability to ignore concurrent speech is more dependent on inhibition than on switching of attention.

4.1 Age-Related Decline in Inhibition of Irrelevant Information and “General Slowing”

In a study of age-related inhibition of irrelevant speech by Li et al. (2004), the ability of young and older participants to inhibit a masker’s speech was tested in a binaural setting with two source positions and a shadowing task involving meaningless sentences. As the older adults were not found to have more difficulty inhibiting the irrelevant information masker in this examination, the results of Li and colleagues opposed the inhibitory deficit theory. It must be noted, however, that the differences in results may be due to the disparate complexities of the binaural conditions. For instance, the setup used by Li and colleagues was simpler compared to the one employed in this study, which involved eight sources around the listener. The difference in source setups may also help explain the disparity between the results of these two studies. In the present study, the congruency effect interacted with the effect of the target’s spatial position, indicating that the congruency effect was highest for the target positioned on the median plane and smallest for the one positioned to the right or the left of the participant. The interaction with age showed significantly higher reaction time differences between source positions for older compared to younger participants. The fact that the older participants were significantly more distracted by the opposing speaker in target positions on the median plane compared to positions on the diagonal plane or to the sides could not be explained adequately. It may be assumed, however, that the applicability of the inhibitory deficit theory is confined to dichotic or elementary spatial listening test conditions, which, for the most part, have been used to support this theory (Braver and Barch 2002; Hasher et al. 2001; Lawo and Koch 2014b). For instance, the congruency effect was found to be least pronounced on the frontal plane (left and right), a situation most comparable to an elementary spatial setup.

In the study by Oberem et al. (2014), young participants showed significantly worse performance when performing the task of ignoring the distractors’ speech, as was evident in the congruency effect when binaural stimuli were non-individual. Therefore, it may be assumed that older people have greater difficulty attending to a target speaker while ignoring the opposing speaker with non-individual binaural stimuli. It is conceivable that older adults suffer more from the loss of individual binaural information, which happens to be particularly crucial for sources located on

the median plane or for competing sources in one cone of confusion (Blauert 1997). The congruency effects seen between the target positions on the median plane and other positions around the listener reinforce this hypothesis. The non-individual HRTFs were measured with an artificial head modeled on the image of a young person (Minnaar et al. 2001), and, therefore, the data might have a better match for young participants as size and shape were age-related (Otte et al. 2013). Further studies would be needed to clarify this issue.

The median plane proved to be most difficult for both young and old participants, with the effect even more pronounced for the latter group. In their localization experiment with participants belonging to different age groups, Abel et al. (2000) also observed greater difficulties with respect to sources positioned in front and back compared to the lateral source positions. The localization performance from younger (10–39 years) to older (60–81 years) participants dropped about 8% for lateral positions and about 12.5% for positions on or close to the median plane. In summary, the task of focusing attention on sources positioned on the median plane was found to be most difficult, with an age-related effect on the horizontal plane compared to other source positions.

The notion of a “general slowing” confirmed in a meta-analysis by Wasylyshyn et al. (2011) using visual tasks has already been corroborated in relation to auditory tasks through the authors’ earlier findings (Lawo and Koch 2014b). The present results are in line with these findings, given the non-significant differences observed in attention switch costs across age groups. The spatial and therefore more complex arrangement of the target’s locations did not influence the participants’ ability to switch attention. Congruency effects involving the categorization of number words were found to increase for older participants contrary to previous findings in a dichotic presentation of auditory stimuli. Thus, the current results in terms of congruency effects appear to be in line with the hypothesis of inhibitory process deficits in older adults (Braver and Barch, 2002; Hasher et al., 2001). Furthermore, the age-related congruency effect was found to depend on the spatial position of the target speaker, with the effect of the deteriorated performance vis-a-vis the median plane compared to other positions on the horizontal plane (Abel et al. 2000; Møller et al. 1996) proving significant in age-related congruency effects.

5 Influence of Reverberation and Head-Movements on Intentional Switching of Auditory Attention in the Extended Binaural Paradigm

Until now, all the experiments described in this chapter are far removed from a real-life scenario due to the anechoic condition. In real-life situations, especially in indoor settings, room acoustics plays an important role, with the reverberation time being an efficient means of characterizing different indoor scenarios. Reverberant energy distorts the signal and increases reaction times and error rates (Nábělek and Robinson

1982; Darwin and Hukin 2000a; Lavandier and Culling 2008), and, therefore, it is of interest in experiments involving auditory selective attention.

In an attention task, Ruggles and Shinn-Cunningham (2011) varied the amount of reverberant energy in three steps: no reverberation ($RT_{60} = 0$ s), low reverberation ($RT_{60} = 0.4$ s), and high reverberation ($RT_{60} = 3$ s). The participants in this study were asked to repeat four consecutive digits. The target speaker was always positioned in front with two other distracting speakers located to the sides. It was found that the reverberation time interfered with selective spatial attention. Similar influences were observed by Culling et al. (2003) and Lavandier and Culling (2007), who found Speech Reception Thresholds (SRTs) to be significantly lower under anechoic conditions. In addition, reverberant energy interacted with the location of target and distractor, indicating no improvement in SRT for spatially separated speakers in the reverberant condition. Kidd et al. (2005b), on the other hand, found the effect of reverberation to be greater when target and masker were spatially separated, rather than being collocated at the same position.

5.1 *Extension of the Binaural Paradigm on Auditory Selective Attention*

The paradigm used in this study was once more further extended to study room acoustic effects (cf., Oberem et al. 2018; Fels et al. 2016). In the previous versions of the paradigm, the stimuli were 700 ms in duration, which is too short to study the influence of reverberation and its interaction with indoor properties. In addition, longer stimuli are more conducive to studying the influence of dynamic head movement. To extend the stimulus duration to 1200 ms, a direction word—“*up*” (German: “*oben*”) and “*down*” (German: “*unten*”)—was added to the digits (e.g., target speaker said: “*three up*” while the distracting speaker articulated: “*seven down*”).

The congruency effect had to be redefined for the analysis of the responses, with the categorization task now containing four response possibilities. Participants were asked to use four buttons on a game-pad for responding. The categories smaller or larger than 5 were mapped to the left- and right-hand buttons of a controller, with the direction words, “*up*” and “*down*” to be categorized using the index and middle fingers of either hand.

Investigating the new paradigm and comparing its results with those of the former binaural paradigm, Fels et al. (2016) found the reaction times and error rates to be generally higher than those seen in the former paradigm (reaction times increased by approx. 100 ms and error rates increased by approx. 3%). This increase in reaction time and error rates may be attributed to the increased number of response alternatives and thus increased the difficulty of response selection of this paradigm (four vs. two response alternatives). The new version of the paradigm is robust and is capable of reproducing findings that are comparable to the ones elicited by the original version.

This extended binaural paradigm was applied in three different rooms with increasing reverberation time. For a room model (total volume of 137 m^3 with a quadrangular ground area) with non-parallel walls, three reverberation conditions were simulated including the anechoic case ($RT_{60} = 0\text{ s}$), a case with low reverberation time ($RT_{60} = 0.8\text{ s}$, comparable to an acoustically untreated classroom), as well as a case with a high reverberation time ($RT_{60} = 1.75\text{ s}$, comparable to an auditorium).

The listener position was placed off-center of the room to prevent unwanted acoustic effects such as room modes (Hartmann 1983; Rakerd and Hartmann 1985; Giguère and Abel 1993). The absorption coefficients in the room model were varied to achieve three levels of reverberation. Binaural room impulse responses (BRIR) were calculated with the software package RAVEN, based on the simulated room model as well as HRTFs of an artificial head measured in an anechoic chamber (Schröder 2012). The dummy head is a mannequin produced at the Institute of Technical Acoustics, RWTH Aachen University, with a simple torso and a detailed ear geometry (Schmitz 1995; Minnaar 2001).

Our experiment yielded effects similar to those found by Kidd et al. (2005a), and Darwin and Hukin (2000b). Reverberation time has been found to have a detrimental effect on reaction time when attention is required to remain focused on one source at a constant spatial location. Intentionally switching attention to a sound source at a different spatial location requires much attention. Additional reverberant energy does not have any further impact on the attention task (Fig. 10). Furthermore, the human ability to ignore or avoid processing the content of a distracting source is influenced significantly by reverberation.

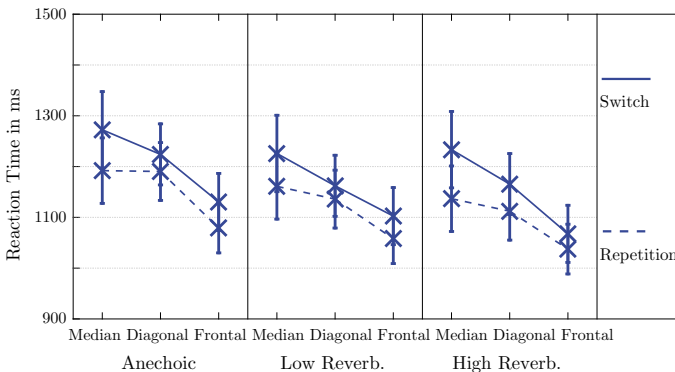


Fig. 10 Reaction time (in ms) as a function of reverberation time, the target's spatial position, and attention switch. Error bars indicate standard errors. The main effect of reverberation was not significant. The main effect of the target's position is shown on the x-axis, describing the target's speaker position in space on the median, diagonal, or frontal plane. The main effect of attention switch can be seen in solid and dotted lines. Older and younger participants react significantly faster when the target's position is repeated and not switched. For more detailed information, see Oberem et al. (2018)

6 Conclusion and Future Directions

This chapter aimed to analyze the influences of variables that increase the complexity of the auditory scene concerning reproduction methods on cognitive control mechanisms underlying auditory selective attention. By comparing performance measures (i.e., reaction time and error rates), reproduction methods (e.g., individual HRTFs and non-individual HRTFs) are validated, and room-acoustical influences (e.g., anechoic, low reverberation time, and high reverberation time) are examined.

The use of dichotic reproduction enabled the observation that participants can easily follow an instruction to switch auditory attention to a new auditory target. However, this switching of attention results in performance costs in terms of increased reaction time and reduced response accuracy. Even though participants succeed in responding to a new auditory target and thus listen selectively to the relevant information, they cannot avoid processing the irrelevant information, to an extent that it can influence their response.

In addition, the dichotic listening paradigm was extended to a binaural listening paradigm which shows great potential for successfully analyzing intentional switching of auditory attention. The extent of individualization of the binaural reproduction method cannot be neglected in terms of absolute values of reaction time and error rates, but the effects on switching auditory attention are not influenced appreciably in an anechoic environment.

The binaural paradigm was also tested in simulated rooms with different reverberation times, because an anechoic reproduction fails to represent realistic listening experiences. It was found that reverberation has a detrimental effect on reaction time when attention is required to be focused on one source at a constant spatial location. Intentionally switching attention to a sound source at a different spatial location appears to require so much more attention that additional reverberant energy does not have a further impact. There may be other variables, such as head and body movements as well as additional distracting noise sources—e.g., cars, airplanes, barking dogs, construction noises, etc.—that contribute to the complexity of an auditory scene, influencing the efficiency of cognitive processing. It is essential to consider both room acoustics and distracting sources when analyzing a natural acoustic scene or creating a dynamic reproduction of an acoustic scenario. Further extensions of the binaural paradigm are therefore necessary in order to examine auditory selective attention in realistic, complex environments.

Auditory attention and processing capacities appear to be contingent on age, with the results of this study in terms of congruency effects being consistent with the hypothesis that older adults suffer from inhibitory deficits. The question of whether there is an age-related decline of auditory attention in complex environments deserves closer inspection. Finally, experiments involving children and hearing-impaired individuals, given their different abilities and challenges, are likely to provide further insight into auditory selective attention in simple and complex acoustic environments.

Acknowledgements The authors gratefully acknowledge the funding provided by the DFG (Deutsche Forschungsgemeinschaft, FE1168/1-1,2 and KO2045/11-1,2). The authors would like to thank Steven van de Par and two anonymous reviewers for their valuable comments and suggestions.

References

- Abel, S.M., C. Giguère, A. Consoli, and B.C. Papsin. 2000. The effect of aging on horizontal plane sound localization. *The Journal of the Acoustical Society of America* 108 (2): 743–752. <https://doi.org/10.1121/1.429607>.
- Allen, K., D. Alais, and S. Carlile. 2009. Speech intelligibility reduces over distance from an attended location: Evidence for an auditory spatial gradient of attention. *Perception & Psychophysics* 71 (1): 164–173. <https://doi.org/10.1121/1.2407738>.
- Bai, M.R., and C.-C. Lee. 2006. Objective and subjective analysis of effects of listening angle on crosstalk cancellation in spatial sound reproduction. *The Journal of the Acoustical Society of America* 120 (4): 1976–1989. <https://doi.org/10.1121/1.2257986>.
- Best, V., F.J. Gallun, A. Ihlefeld, and B.G. Shinn-Cunningham. 2006. The influence of spatial separation on divided listening. *The Journal of the Acoustical Society of America* 120: 1506. <https://doi.org/10.1121/1.2234849>.
- Best, V., E.J. Ozmeral, F.J. Gallun, K. Sen, and B.G. Shinn-Cunningham. 2005. Spatial unmasking of birdsong in human listeners: Energetic and informational factors. *The Journal of the Acoustical Society of America* 118: 3766. <https://doi.org/10.1121/1.2130949>.
- Best, V., E.J. Ozmeral, and B.G. Shinn-Cunningham. 2007. Visually-guided attention enhances target identification in a complex auditory scene. *JARO—Journal of the Association for Research in Otolaryngology* 8 (2): 294–304. <https://doi.org/10.1007/s10162-007-0073-z>.
- Best, V., B.G. Shinn-Cunningham, E.J. Ozmeral, and N. Kopčo. 2010. Exploring the benefit of auditory spatial continuity. *The Journal of the Acoustical Society of America* 127 (6): EL258–EL264. <https://doi.org/10.1121/1.3431093>.
- Blauert, J. 1997. *Spatial Hearing—The Psychophysics of Human Sound Localization*, 2nd ed. Cambridge MA: MIT Press.
- Blauert, J., and J. Braasch. 2008. Räumliches Hören [Spatial hearing]. In *Handbuch der Audiotechnik*, ed. S. Weinzierl, 87–122. Berlin/Heidelberg: Springer. <https://doi.org/10.1002/bapi.200890058>.
- Braver, T.S., and D.M. Barch. 2002. A theory of cognitive control, aging cognition, and neuro-modulation. *Neuroscience & Biobehavioral Reviews* 26 (7): 809–817. [https://doi.org/10.1016/S0149-7634\(02\)00067-2](https://doi.org/10.1016/S0149-7634(02)00067-2).
- Broadbent, D.E. 1958. Effect of noise on an “intellectual” task. *The Journal of the Acoustical Society of America* 30 (9): 824–827.
- Bronkhorst, A.W. 1995. Localization of real and virtual sound sources. *Journal of the Acoustical Society of America* 98 (5): 2542–2553. <https://doi.org/10.1121/1.413219>.
- Bronkhorst, A.W. 2015. The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics* 77 (5): 1465–1487. <https://doi.org/10.3758/s13414-015-0882-9>.
- Brungart, D.S., and B.D. Simpson. 2004. Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty. *The Journal of the Acoustical Society of America* 115 (1): 301–310. <https://doi.org/10.1121/1.1628683>.
- Butler, R.A., and K. Belendiuk. 1977. Spectral cues utilized in the localization of sound in the median sagittal plane. *Journal of the Acoustical Society of America* 61: 1264–1269. <https://doi.org/10.1121/1.381427>.
- Cherry, E.C. 1953. Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America* 25 (5): 975–979. <https://doi.org/10.1121/1.1907229>.

- Culling, J.F., K.I. Hodder, and C.Y. Toh. 2003. Effects of reverberation on perceptual segregation of competing voices. *The Journal of the Acoustical Society of America* 114 (5): 2871–2876. <https://doi.org/10.1121/1.1616922>.
- Darwin, C.J., and R.W. Hukin. 2000a. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of America* 107 (2): 970–977. <https://doi.org/10.1121/1.428278>.
- Darwin, C.J., and R.W. Hukin. 2000b. Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. *The Journal of the Acoustical Society of America* 108 (1): 335–342. <https://doi.org/10.1121/1.429468>.
- Dobrevá, M.S., W.E. O'Neill, and G.D. Paige. 2011. Influence of aging on human sound localization. *Journal of Neurophysiology* 105 (5): 2471–2486. <https://doi.org/10.1152/jn.00951.2010>.
- Duquesnoy, A.J. 1983. The intelligibility of sentences in quiet and in noise in aged listeners. *The Journal of the Acoustical Society of America* 74 (4): 1136–1144. <https://doi.org/10.1121/1.390037>.
- Fels, J., J. Oberem, and I. Koch. 2016. Examining auditory selective attention in realistic, natural environments with an optimized paradigm. *Proceedings of Meetings on Acoustics* 28 (1): 050001. <https://doi.org/10.1121/2.0000321>.
- Freedman, S., and H. Fisher. 1968. *Neuropsychology of Spatially Oriented Behaviour: The Role of the Pinna in Auditory Localization*. Homewood, Illinois: Dorsey Press.
- Gardner, W.G. 1997. *3-d Audio Using Loudspeakers*. Massachusetts USA: Massachusetts Institute of Technology. Ph.D. thesis.
- Getzmann, S., C. Hanenberg, J. Lewald, M. Falkenstein, and E. Wascher. 2015. Effects of age on electrophysiological correlates of speech processing in a dynamic “cocktail-party” situation. *Frontiers in Neuroscience* 9: 341. <https://doi.org/10.3389/fnins.2015.00341>.
- Giguère, C., and S.M. Abel. 1993. Sound localization: Effects of reverberation time, speaker array, stimulus frequency, and stimulus rise/decay. *The Journal of the Acoustical Society of America* 94 (2): 769–776. <https://doi.org/10.1121/1.408206>.
- Hartmann, W.M. 1983. Localization of sound in rooms. *The Journal of the Acoustical Society of America* 74 (5): 1380–1391. <https://doi.org/10.1121/1.390163>.
- Hartmann, W.M., and A. Wittenberg. 1996. On the externalization of sound images. *Journal of the Acoustical Society of America* 99: 3678–3688. <https://doi.org/10.1121/1.414965>.
- Hasher, L., S.T. Tonev, C. Lustig, and R.T. Zacks. 2001. Inhibitory control, environmental support, and self-initiated processing in aging. In *Perspectives on Human Memory and Cognitive Aging: Essays in Honour of Fergus Craik*, ed. M. Naveh-Benjamin, M. Moscovitch, and R.L. Roediger, 286–297. East Sussex, England: Psychology Press.
- Helfer, K.S., C.R. Mason, and C. Marino. 2013. Aging and the perception of temporally-interleaved words. *Ear and Hearing* 34 (2): 160–167. <https://doi.org/10.1097/AUD.0b013e31826a8ea7>.
- Holender, D. 1986. Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioral and Brain Sciences* 9 (1): 1–23. <https://doi.org/10.1017/S0140525X00021269>.
- Hugdahl, K. 2011. Fifty years of dichotic listening research—Still going and going and. . . *Brain and Cognition* 76 (2): 211–213. <https://doi.org/10.1016/j.bandc.2011.03.006>.
- Humes, L.E., J.H. Lee, and M.P. Coughlin. 2006. Auditory measures of selective and divided attention in young and older adults using single-talker competition. *The Journal of the Acoustical Society of America* 120 (5): 2926–2937. <https://doi.org/10.1121/1.2354070>.
- Iwaya, Y., Y. Suzuki, and D. Kimura. 2003. Effects of head movement on front-back error in sound localization. *Acoustical Science and Technology* 24 (5): 322–324. <https://doi.org/10.1250/ast.24.322>.
- Jongkees, L.B., and D. Veer. 1958. On directional sound localization in unilateral deafness and its explanation. *Acta Oto-Laryngologica* 49 (1): 119–131. <https://doi.org/10.3109/00016485809134735>.
- Jost, K., W. De Baene, I. Koch, and M. Brass. 2013. A review of the role of cue processing in task switching. *Zeitschrift für Psychologie*. <https://doi.org/10.1027/2151-2604/a000125>.

- Kidd, G., T.L. Arbogast, C.R. Mason, and F.J. Gallun. 2005a. The advantage of knowing where to listen. *The Journal of the Acoustical Society of America* 118 (6): 3804–3815. <https://doi.org/10.1121/1.2109187>.
- Kidd, G., C.R. Mason, A. Brughera, and W.M. Hartmann. 2005b. The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta Acustica United with Acustica* 91 (3): 526–536.
- Kiesel, A., M. Steinhauser, M. Wendt, M. Falkenstein, K. Jost, and A.M. Philipp. 2010. Control and interference in task switching—a review. *Psychological Bulletin* 136 (5): 849. <https://doi.org/10.1037/a0019842>.
- Kitterick, P.T., P.J. Bailey, and A.Q. Summerfield. 2010. Benefits of knowing who, where, and when in multi-talker listening. *The Journal of the Acoustical Society of America* 127 (4): 2498–2508. <https://doi.org/10.1121/1.3327507>.
- Koch, I., M. Gade, S. Schuch, and A. Philipp. 2010. The role of inhibition in task switching: A review. *Psychonomic Bulletin & Review* 17 (1): 1–14. <https://doi.org/10.3758/PBR.17.1.1>.
- Koch, I., and V. Lawo. 2014. Exploring temporal dissipation of attention settings in auditory task switching. *Attention, Perception, & Psychophysics* 76: 73–80. <https://doi.org/10.3758/s13414-013-0571-5>.
- Koch, I., V. Lawo, J. Fels, and M. Vorländer. 2011. Switching in the cocktail party: Exploring intentional control of auditory selective attention. *Journal of Experimental Psychology/Human Perception and Performance* 37 (4): 1140–1147. <https://doi.org/10.1037/a0022189>.
- Koch, I., E. Poljac, H. Müller, and A. Kiesel. 2018. Cognitive structure, flexibility, and plasticity in human multitasking—an integrative review of dual-task and task-switching research. *Psychological Bulletin* 144 (6): 557. <https://doi.org/10.1037/bul0000144>.
- Kramer, A.F., S. Hahn, and D. Gopher. 1999. Task coordination and aging: explorations of executive control processes in the task switching paradigm. *Acta Psychologica* 101 (2–3): 339–378. [https://doi.org/10.1016/S0001-6918\(99\)00011-6](https://doi.org/10.1016/S0001-6918(99)00011-6).
- Kray, J., J. Eber, and J. Karbach. 2008. Verbal self-instructions in task switching: A compensatory tool for action-control deficits in childhood and old age? *Developmental Science* 11 (2): 223–236. <https://doi.org/10.1111/j.1467-7687.2008.00673.x>.
- Lachter, J., K.I. Forster, and E. Ruthruff. 2004. Forty-five years after broadbent (1958): Still no identification without attention. *Psychological Review* 114 (4): 880. <https://doi.org/10.1037/0033-295X.111.4.880>.
- Langendijk, E.H.A., and A.W. Bronkhorst. 2000. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *The Journal of the Acoustical Society of America* 107 (1): 528–537. <https://doi.org/10.1121/1.428321>.
- Lavandier, M., and J.F. Culling. 2007. Speech segregation in rooms: Effects of reverberation on both target and interferer. *The Journal of the Acoustical Society of America* 122 (3): 1713–1723. <https://doi.org/10.1121/1.2764469>.
- Lavandier, M., and J.F. Culling. 2008. Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer. *The Journal of the Acoustical Society of America* 123 (4): 2237–2248. <https://doi.org/10.1121/1.2871943>.
- Lavie, N. 2005. Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences* 9 (2): 75–82. <https://doi.org/10.1016/j.tics.2004.12.004>.
- Lawo, V., J. Fels, J. Oberem, and I. Koch. 2014. Intentional attention switching in dichotic listening: Exploring the efficiency of nonspatial and spatial selection. *The Quarterly Journal of Experimental Psychology* 67 (10): 2010–2024. <https://doi.org/10.1080/17470218.2014.898079>.
- Lawo, V., and I. Koch. 2014a. Dissociable effects of auditory attention switching and stimulus-response compatibility. *Psychological Research* 78 (3): 379–386. <https://doi.org/10.1007/s00426-014-0545-9>.
- Lawo, V., and I. Koch. 2014b. Examining age-related differences in auditory attention control using a task-switching procedure. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 69: 237–244. <https://doi.org/10.1093/geronb/gbs107>.

- Lawo, V., and I. Koch. 2015. Attention and action: The role of response mappings in auditory attention switching. *Journal of Cognitive Psychology* 27 (2): 194–206. <https://doi.org/10.1080/20445911.2014.995669>.
- Lentz, T., I. Assenmacher, and J. Sokoll. 2005. Performance of spatial audio using dynamic cross-talk cancellation. In *119th Audio Engineering Society Convention*, New York, NY, US, Preprint No. 6541, Permalink: <http://www.aes.org/e-lib/browse.cfm?elib=13296>.
- Li, L., M. Daneman, J.G. Qi, and B.A. Schneider. 2004. Does the information content of an irrelevant source differentially affect spoken word recognition in younger and older adults? *Journal of Experimental Psychology: Human Perception and Performance* 30 (6): 1077–1091. <https://doi.org/10.1037/0096-1523.30.6.1077>.
- Lindau, A., and S. Weinzierl. 2012. Assessing the plausibility of virtual acoustic environments. *Acta Acustica United with Acustica* 98 (5): 804–810. <https://doi.org/10.3813/AAA.918562>.
- Logan, G.D. 2005. The time it takes to switch attention. *Psychonomic Bulletin & Review* 12 (4): 647–653. <https://doi.org/10.3758/BF03196753>.
- Logan, G.D., and R.D. Gordon. 2001. Executive control of visual attention in dual-task situations. *Psychological Review* 108 (2): 393.
- Marrone, N., C.R. Mason, and G. Kidd. 2008. The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *The Journal of the Acoustical Society of America* 124 (5): 3064–3075. <https://doi.org/10.1121/1.2980441>.
- Masiero, B., and J. Fels. 2011. Perceptually robust headphone equalization for binaural reproduction. In *130th Audio Engineering Society Convention*, New York, NY, US, p. 8388
- Meiran, N. 1996. Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22 (6): 1423. <https://doi.org/10.1037/0278-7393.22.6.1423>.
- Minnaar, P. 2001. Simulating an acoustical environment with binaural technology—investigations of binaural recording and synthesis. Ph.D. thesis, Denmark: Aalborg University, Aalborg.
- Minnaar, P., S.K. Olesen, F. Christensen, and H. Møller. 2001. Localization with binaural recordings from artificial and human heads. *Journal of Audio Engineering Society* 49: 323–336.
- Møller, H., C.B. Jensen, D. Hammershøi, and M.F. Sørensen. 1996. Using a typical human subject for binaural recording. In *100th Audio Engineering Society Convention*, Preprint No. 4157, Permalink: <http://www.aes.org/e-lib/browse.cfm?elib=7614>.
- Mondor, T.A., R.J. Zatorre, and N.A. Terrio. 1998. Constraints on the selection of auditory information. *Journal of Experimental Psychology: Human Perception and Performance* 24 (1): 66.
- Monsell, S. 2003. Task switching. *Trends in Cognitive Sciences* 7 (3): 134–140. [https://doi.org/10.1016/S1364-6613\(03\)00028-7](https://doi.org/10.1016/S1364-6613(03)00028-7).
- Moore, A.H., A.I. Tew, and R. Nicol. 2010. An initial validation of individualized crosstalk cancellation filters for binaural perceptual experiments. *Journal of Audio Engineering Society* 58 (1/2): 36–45.
- Nábělek, A.K., and P.K. Robinson. 1982. Monaural and binaural speech perception in reverberation for listeners of various ages. *The Journal of the Acoustical Society of America* 71 (5): 1242–1248. <https://doi.org/10.1121/1.387773>.
- Oberem, J., I. Koch, and J. Fels. 2017. Intentional switching in auditory selective attention: Exploring age-related effects in a spatial setup requiring speech perception. *Acta Psychologica* 177: 36–43. <https://doi.org/10.1016/j.actpsy.2017.04.008>.
- Oberem, J., V. Lawo, I. Koch, and J. Fels. 2014. Intentional switching in auditory selective attention: Exploring different binaural reproduction methods in an anechoic chamber. *Acta Acustica United with Acustica* 100 (6): 1139–1148. <https://doi.org/10.3813/AAA.918793>.
- Oberem, J., B. Masiero, and J. Fels. 2016. Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods. *Applied Acoustics* 114: 71–78. <https://doi.org/10.1016/j.apacoust.2016.07.009>.
- Oberem, J., J. Seibold, I. Koch, and J. Fels. 2018. Intentional switching in auditory selective attention: Exploring attention shifts with different reverberation times. *Hearing Research* 359: 32–39. <https://doi.org/10.1016/j.heares.2017.12.013>.

- Otte, R.J., M.J.H. Agterberg, M.M. Wanrooij, A.F.M. Snik, and A.J. Opstal. 2013. Age-related hearing loss and ear morphology affect vertical but not horizontal sound-localization performance. *Journal of the Association for Research in Otolaryngology* 14 (2): 261–273. <https://doi.org/10.1007/s10162-012-0367-7>.
- Pashler, H.E. 1998. *The Psychology of Attention*, vol. 15. Cambridge, MA: MIT Press.
- Perrett, S., and W. Noble. 1997a. The contribution of head motion cues to localization of low-pass noise. *Attention, Perception & Psychophysics* 59: 1018–1026. <https://doi.org/10.3758/BF03205517>.
- Perrett, S., and W. Noble. 1997b. The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America* 102 (4): 2325–2332. <https://doi.org/10.1121/1.419642>.
- Rakerd, B., and W.M. Hartmann. 1985. Localization of sound in Rooms, II: The effects of a single reflecting surface. *The Journal of the Acoustical Society of America* 78 (2): 524–533. <https://doi.org/10.1121/1.392474>.
- Rivenez, M., A. Guillaume, L. Bourgeon, and C.J. Darwin. 2008. Effect of voice characteristics on the attended and unattended processing of two concurrent messages. *European Journal of Cognitive Psychology* 20 (6): 967–993. <https://doi.org/10.1080/09541440701686201>.
- Ruggles, D., and B.G. Shinn-Cunningham. 2011. Spatial selective auditory attention in the presence of reverberant energy: Individual differences in normal-hearing listeners. *Journal of the Association for Research in Otolaryngology* 12: 395–405. <https://doi.org/10.1007/s101620100254z>.
- Schärer, Z., and A. Lindau. 2009. Evaluation of equalization methods for binaural signals. In *126th Audio Engineering Society Convention*, New York, NY, US, Preprint No. 7721, Permalink: <http://www.aes.org/e-lib/browse.cfm?elib=14917>.
- Schmitz, A. 1995. Ein neues digitales Kunstkopfmesssystem (A new digital dummy-head measurement system). *Acta Acustica United with Acustica* 81 (4): 416–420.
- Schröder, D. (2012). Physically based real-time Auralization of interactive virtual environments. Dissertation, RWTH Aachen University, Aachen Germany.
- Searle, C., L. Braida, D. Cuddy, and M. Davis. 1975. Binaural pinna disparity: Another auditory localization cue. *The Journal of the Acoustical Society of America* 57: 448–455. <https://doi.org/10.1121/1.380442>.
- Seibold, J.C., S. Nolden, J. Oberem, J. Fels, and I. Koch. 2018. Intentional preparation of auditory attention-shifts: Explicit cueing and sequential shift-predictability. *The Quarterly Journal of Experimental Psychology* 71: 1382–1395. <https://doi.org/10.1080/17470218.2017.1344867>.
- Shinn-Cunningham, B.G. 2008. Object-based auditory and visual attention. *Trends in Cognitive Sciences* 12 (5): 182–186. <https://doi.org/10.1016/j.tics.2008.02.003>.
- Singh, G., M.K. Pichora-Fuller, and B.A. Schneider. 2013. Time course and cost of misdirecting auditory spatial attention in younger and older adults. *Ear and Hearing* 34 (6): 711–721. <https://doi.org/10.1097/AUD.0b013e31829bf6ec>.
- Takeuchi, T., P.A. Nelson, and H. Hamada. 2001. Robustness to head misalignment of virtual sound imaging systems. *The Journal of the Acoustical Society of America* 109 (3): 958–971. <https://doi.org/10.1121/1.1349539>.
- Thurlow, W.R., J.W. Mangels, and P.S. Runge. 1967. Head movements during sound localization. *The Journal of the Acoustical Society of America* 42 (2): 489–493. <https://doi.org/10.1121/1.1910605>.
- Toshima, I., and Aoki, S. 2006. The effect of head movement on sound localization in an acoustical telepresence robot: Telehead. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, <https://doi.org/10.1109/IROS.2006.281740>.
- Treisman, A.M. 1969. Strategies and models of selective attention. *Psychological Review* 76 (3): 282. <https://doi.org/10.1037/h0027242>.
- Tun, P.A., G. O’Kane, and A. Wingfield. 2002. Distraction by competing speech in young and older adult listeners. *Psychology and Aging* 17 (3): 453–467. <https://doi.org/10.1037/0882-7974.17.3.453>.

- Vandierendonck, A., B. Liefoghe, and F. Verbruggen. 2010. Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin* 136 (4): 601–626. <https://doi.org/10.1037/a0019791>.
- Wallach, H. 1940. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology* 27: 339–368. <https://doi.org/10.1037/h0054629>.
- Wasylyshyn, C., P. Verhaeghen, and M.J. Sliwinski. 2011. Aging and task switching: A meta-analysis. *Psychology and Aging* 26 (1): 15. <https://doi.org/10.1037/a0020912>.
- Wenzel, E.M., M. Arruda, D.J. Kistler, and F.L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94 (1): 111–123. <https://doi.org/10.1121/1.407089>.
- Wightman, F.L., and D.J. Kistler. 1989. Headphone simulation of free-field listening. ii: Psychophysical validation. *The Journal of the Acoustical Society of America* 85 (2): 868–878. <https://doi.org/10.1121/1.397558>.
- Wood, N., and N. Cowan. 1995a. The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21 (1): 255. <https://doi.org/10.1037/0278-7393.21.1.255>.
- Wood, N.L., and N. Cowan. 1995b. The cocktail party phenomenon revisited: Attention and memory in the classic selective listening procedure of Cherry (1953). *Journal of Experimental Psychology: General* 124 (3): 243. <https://doi.org/10.1037/0096-3445.124.3.243>.
- Woods, D.L., C. Alain, R. Diaz, D. Rhodes, and K.H. Ogawa. 2001. Location and frequency cues in auditory selective attention. *Journal of Experimental Psychology: Human Perception and Performance* 27 (1): 65.
- Young, P.T. 1931. The role of head movements in auditory localization. *Journal of Experimental Psychology* 14 (2): 95. <https://doi.org/10.1037/h0075721>.
- Zahorik, P., F.L. Wightman, and D.J. Kistler. 1996. The fidelity of virtual auditory displays. *The Journal of the Acoustical Society of America* 99 (4): 2596. <https://doi.org/10.1121/1.415284>.

Blackboard Systems for Cognitive Audition



Christopher Schymura and Dorothea Kolossa

Abstract An essential part of auditory scene understanding is building an internal model of the world surrounding the listener. This internal representation can be mimicked computationally via a blackboard-system-based software architecture. Blackboard systems allow efficient integration of different perceptual modalities, algorithms, and data representations into a coherent and flexible computational framework. The term “blackboard” in this context stands for a flexible and compositional internal data representation, allowing individual software modules to access and process available information. This modular architecture also makes the system adaptable to different application scenarios and provides interfaces to incorporate feedback paths, which allows the system to derive task-optimal active behavior from the internal model. Extending conventional blackboard systems with modern machine-learning techniques, specifically probabilistic modeling and neural networks, enables the system to incorporate learning strategies into this computational framework. Additionally, online learning and adaptation strategies can be integrated into the data representation within the blackboard. This is particularly useful for developing feedback approaches. This chapter gives a review of existing blackboard systems for different applications and provides the necessary theoretical foundations. Subsequently, novel extensions that were recently introduced in the context of binaural scene analysis and understanding are presented and discussed. A special focus is set on possibilities for incorporating feedback and learning strategies into the framework.

C. Schymura (✉) · D. Kolossa
Institute of Communication Acoustics, Faculty of Electrical Engineering
and Information Technology, Ruhr-Universität Bochum, 44801 Bochum, Germany
e-mail: christopher.schymura@rub.de

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_4

1 Introduction

Human listeners have a remarkable ability to assess complex acoustic scenes, even under adverse conditions involving background noise and reverberation. This phenomenon has been termed *auditory scene analysis* (ASA) by Bregman (1990) and has since been extensively investigated with a focus on reproducing this ability by computational models—see, for example, Wang and Brown (2006). In order to do so, machine-hearing systems may achieve a perceptual organization of sound in a similar way as human listeners do. This involves the integration of diverse sources of knowledge, including primitive grouping heuristics, as well as schema-driven grouping principles. Additionally, the interaction between bottom-up and top-down processes through feedback loops plays an important role which allows the system to adapt depending on higher-level control tasks.

Recent approaches to meet these requirements for a machine-hearing system are based on blackboard problem-solving architectures, as for instance proposed by Schymura et al. (2014). A blackboard system consists of a group of independent experts called *knowledge sources* (KS), which communicate by reading and writing data on a globally-accessible data structure—the blackboard. Typically, the blackboard is divided into layers that correspond to low-level sensory data, processed auditory features, hypotheses and partial solutions at different levels of abstraction. A third component, the scheduler, coordinates actions that individual knowledge sources can perform, based on the current state of the blackboard.

Blackboard systems were initially proposed by Erman et al. (1980) in the context of speech understanding and have since been applied to a variety of problems in different subject areas. Hayes-Roth (1985) introduced a general blackboard architecture to solve planning and control tasks. Based on this initial line of research, blackboard systems for specific technical applications have emerged. These include, for instance, a framework for real-time mobile robot navigation proposed by Pang (1988) and an expert system for controlling structure synthesis in chemical-processing plants—see Song et al. (1991). Besides the domain of classical control, early research on blackboard systems also contributed to other technical fields. Exemplary applications were technical diagnosis (Hong et al. 1997), the design of electrical components (Dirand and Chevrier 1995) and knowledge systems (Hewett and Hewett 1993). Besides the deployment in specific technical applications, the classical blackboard model of problem-solving, as proposed by Erman et al. (1980), was subsequently refined and extended. Weiss and Stetter (1992) proposed a hierarchical blackboard architecture, specifically designed to solve problems that can be effectively decomposed into individual subtasks. Furthermore, an initial formal description of the blackboard problem-solving model was introduced by McManus and Bynum (1996). Herein, the authors also present a set of tools for design, the simulation, and refinement of blackboard architectures. Recent developments include multi-agent and information-fusion systems (Hou et al. 2000; Zhu et al. 2010), knowledge and information management (Barot et al. 2013; Rong 2014) and distributed systems like mobile sensor networks (Wang et al. 2011).

The blackboard architecture has specific characteristics that make it especially suitable for machine hearing. Among other features, it provides a framework for reasoning about acoustic scenes that is flexible, opportunistic, and integrates bottom-up processing with top-down feedback. This chapter shall serve as an introduction to the general framework of blackboard architectures and their applications. Herein, the focus is set on their applicability towards machine hearing and related problems. After shortly reviewing the historical development of blackboard systems and establishing the theoretical foundations, exemplary architectural blackboard designs suited for specific tasks are introduced and discussed. This includes conventional architectures similar to the initially proposed systems, as well as recent extensions that combine blackboard systems with modern machine-learning methodologies to solve complex problems. The chapter concludes with a brief summary and provides an outlook on possible research directions involving blackboard systems in the context of machine hearing.

2 Blackboard Systems

Blackboard systems were introduced by Erman et al. (1980) as an architecture for speech understanding, in their Hearsay-II system. Subsequently, a number of authors described blackboard-based systems for machine-hearing applications, for example, Cooke et al. (1993); Lesser et al. (1995); Ellis (1996); Godsmark and Brown (1999). All of these systems were in most respects based on traditional approaches from the area of artificial-intelligence research, especially focusing on rule-based heuristics. Additionally, more recent developments have extended these approaches towards modern machine-learning techniques like, for instance, Bayesian methods as introduced by Sutton et al. (2004). These blackboard systems have been successfully applied in different domains, such as musical-pitch estimation and analysis (Godsill and Davy 2002) and robotics (Fox et al. 2012). In this section, the general architecture and fundamental building blocks of blackboard systems are introduced and put into context according to their historical developments.

2.1 *The Blackboard Model of Problem Solving*

The basic idea behind using blackboard systems for solving complex problems is often presented using a metaphor, which is being quoted here from Corkill (1991).

Imagine a group of human specialists seated next to a large blackboard. The specialists are working cooperatively to solve a problem, using the blackboard as the workplace for developing the solution.

Problem-solving begins when the problem and initial data are written onto the blackboard. The specialists watch the blackboard, looking for an opportunity to apply their expertise to the developing solution.

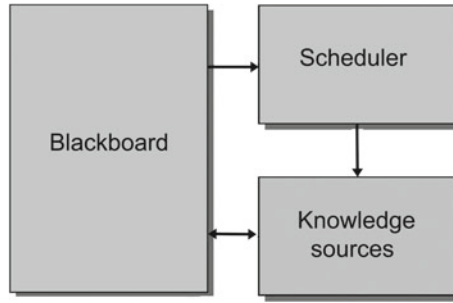


Fig. 1 General architecture of a blackboard system, based on the framework proposed in Corkill (1991). The **arrow directions** indicate data flow. This architecture deviates from the originally proposed idea by Corkill (1991) in the sense that the knowledge sources are enabled to access external data sources which are not associated with the blackboard. The ability to incorporate external data sources introduces additional flexibility to the classical blackboard framework, making it especially suitable to deal with time-series data and real-time processing

When a specialist finds sufficient information to make a contribution, she records the contribution on the blackboard, hopefully enabling other specialists to apply their expertise. This process of adding contributions to the blackboard continues until the problem has been solved

This simple metaphor already captures the basic components and workflow of a computational blackboard system. Following the initially proposed model of Erman et al. (1980), the most general structure possible consists of three components, that is, the blackboard itself, a set of knowledge sources and a scheduler. An overview of the general blackboard architecture as assumed in this chapter is shown in Fig. 1.

The blackboard corresponds to the central data repository of the system, by providing a global database that maintains all input data and partial solutions. It not only stores current data but also keeps track of the history of these data in order to enable working with time series. Like the physical blackboard from the metaphor quoted above, data can be added, removed and modified at any time. In the most general case, data representations can be arbitrary, that is, comprise numerical, probabilistic and/or semantic data. However, specific implementations of blackboard architectures might restrict data representation due to computational or problem-specific restrictions.

Knowledge sources are software modules that define their own functionality, to be executed in the organized frame of the system. They define which data they need for execution and which data they produce. The blackboard system provides the tools for requesting and storing these data but does not care about the actual contents, while the knowledge sources need not care about where and how the data are stored. These modules can work on different levels of abstraction, independently from each other or in collaboration, in a bottom-up or in a top-down manner. From an abstract viewpoint, knowledge sources can be seen as specialist or expert modules, similar to human experts. Generally, each knowledge source is designed to solve a specific subtask that contributes to the solution of the addressed problem at large.

The scheduler is the component of the blackboard system that actually executes the knowledge source. Most importantly, it determines the order in which knowledge sources get executed, based on the current task, the data that are stored on the blackboard, and of available computational resources. This order is rescheduled after every execution of a knowledge source, since the conditions determining the order may have changed, or new knowledge sources may be waiting for execution.

Despite the three basic components, the specific design and implementation of a blackboard system always depends on the problem to be solved. Possible extensions of the conventional structure discussed here are introduced in Sect. 4.

2.2 Blackboard Systems as Computational Frameworks

To implement a blackboard system within a computational framework, a more formal definition of the concepts introduced in Sect. 2.1 is required. Therefore, let \mathcal{S} denote the set of all possible blackboard states, which depend on the restrictions imposed on possible data representations. If, for instance, only discrete data representations are allowed to be stored on the blackboard (Nii 1986), \mathcal{S} will comprise a finite-state space. However, when using continuous data representations, \mathcal{S} might also span an infinite state space. By defining $s_i \in \mathcal{S}$ as the actual state of the blackboard at iteration step i and further denoting $x_i \subseteq s_i$ as a subset of the blackboard state serving as input to the m -th KS, the set of knowledge sources can be defined as

$$\mathcal{K} = \left\{ f_1(x_i, y_i, \theta_1), \dots, f_M(x_i, y_i, \theta_M) \right\}, \quad (1)$$

where y_i represents external input data and θ_m are knowledge-source specific parameters. It should be noted that both, y_i and θ_m , are optional and not required for designing a knowledge source. Hence, an individual knowledge source can be interpreted as a mapping $f_m : x_i \rightarrow z_i$, where z_i are new data produced by the m -th KS, which subsequently is appended to the blackboard's state according to $s_{i+1} \leftarrow s_i \cup z_i$. The internal parameterization, θ_m , introduces additional flexibility to the design of the knowledge sources. This allows, for example, for the training of individual knowledge sources by using supervised-learning techniques before deploying them on the blackboard framework. An example in the context of auditory-scene understanding is the classification of the acoustic environment that surrounds the listener—based on the audio signals captured. An exemplary implementation of this specific task is depicted and explained in Fig. 2. This example already provides a first impression of the integration of feedback mechanisms with the blackboard framework. This process is further discussed in Sect. 2.3.

Last but not least, the scheduler serves as a control instance that selects knowledge sources according to the current state of the blackboard. The specific implementation of this mechanism may vary depending on the problem that has to be solved. Generally, the scheduler is able to select an action, $a_i \in \mathcal{A}$, from a finite set of pos-

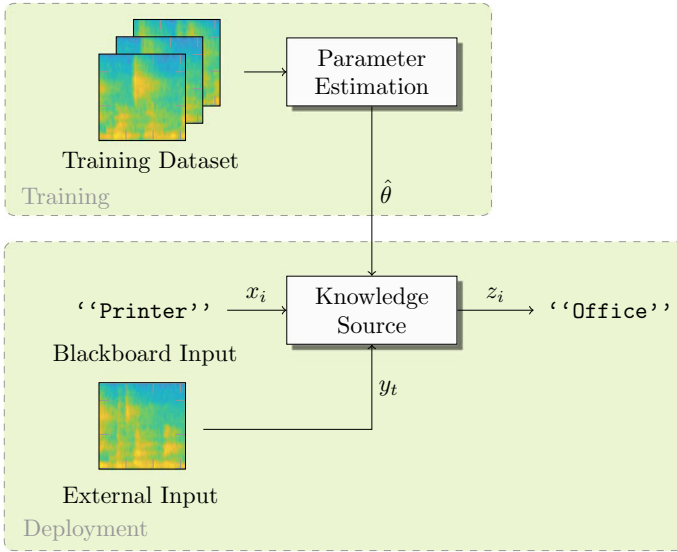


Fig. 2 This example illustrates a knowledge source performing acoustic scene classification based on captured audio data. A dedicated training dataset with labeled examples allows for training parameters, $\hat{\theta}$, for a classifier. The classifier can be used to infer the acoustic scene from external audio data, y_t , as obtained during deployment. A refinement of the resulting hypothesis about the acoustic scene, z_t , can be performed by incorporating additional information from the blackboard, x_t . In the current example, it is assumed that a further knowledge source that is able to detect acoustic events, has put up the hypothesis on the blackboard that the sound of a printer is present in the acoustic signals. This additional information can then be used to prime the currently active scene-classification knowledge source—for instance, by imposing a prior on a probabilistic-classification result. The resulting hypothesis, z_t , is subsequently added to the blackboard data repository and serves as an additional input to other knowledge sources in the next iteration step

sible actions, \mathcal{A} , at each iteration. In the most basic case, an action corresponds to the selection of a knowledge source from the set specified in (1). Depending on the current state of the blackboard, not all possible actions may be available at each iteration—for instance, if certain knowledge sources cannot be executed due to a lack of matching input data on the blackboard. The action-selection process of the scheduler can be expressed via a policy, $\pi(a_i | s_i)$, which might be either stochastic or deterministic. This notion is inspired by reinforcement learning, a technique which is mainly concerned with the learning of optimal policies for control problems—compare Sutton and Barto (1998). To implement a functional blackboard system, at least three different types of actions have to be defined, namely,

- An action that selects a knowledge source based on the available input data, executes it, and writes the resulting new data back to the blackboard. This action is termed `SelectKS` throughout this chapter and comprises two additional parameters, namely, the knowledge-source index, m , and the corresponding data subset, x_i . Both should be read from the blackboard.

Algorithm 1 Canonical blackboard processing framework with time-series data as an external input

Require: A set of knowledge sources \mathcal{K} , a scheduling policy $\pi(a_i | s_i, y_t)$ and a sequence of external input data y_1, \dots, y_T .

```

1:  $s_0 = \emptyset$                                 ▷ Initialize blackboard as an empty set.
2: for  $t = 1$  to  $T$  do
3:    $i = 1$                                     ▷ Initialize blackboard iteration counter.
4:   repeat
5:      $a_i \sim \pi(a_i | s_i, y_t)$               ▷ Sample action from policy.
6:     if  $a_i = \{\text{SelectKS}, m, x_i\}$  then
7:        $z_i = f_m(x_i, y_t, \theta_m)$ 
8:        $s_{i+1} \leftarrow s_i \cup z_i$           ▷ Execute KS and add result to blackboard.
9:     else if  $a_i = \{\text{DeleteData}, x_i\}$  then
10:       $s_{i+1} \leftarrow s_i \cap x_i$         ▷ Remove data from blackboard.
11:      $i = i + 1$ 
12:   until  $a_i = \{\text{Terminate}\}$ 

```

- The ability to remove data from the blackboard that are not required anymore is an important aspect of preventing unnecessary occupation of memory. Therefore, an action `DeleteData` is introduced. This action has one additional parameter, x_i , which describes the subset of data that should be deleted from the blackboard.
- An action, `Terminate`, determines further processing on the blackboard which is not required anymore, hence terminates the processing loop.

Additional actions may be defined during the design of a blackboard system, if necessary or helpful. However, in its most basic form, a functional blackboard system can be built using just these three types of actions described above.

The general concepts of a blackboard, namely, data repository, knowledge sources, and scheduler, as introduced above, allow for expressing the overall blackboard architecture as a computational algorithm as outlined in Algorithm 1. This is a very basic form of a possible blackboard-processing framework, which offers the flexibility to be extended based on the specific requirements of the problem that should be solved by the system. The framework presented here deviates from the original proposal of, for instance, Erman et al. (1980) and Corkill (1991). This is because it explicitly incorporates the concept of a policy for the scheduler and allows knowledge sources to have internal parameters. However, this general framework is already flexible enough to be adaptable to a broad range of application domains. A graphical illustration of the operations that are basically supported is depicted in Fig. 3.

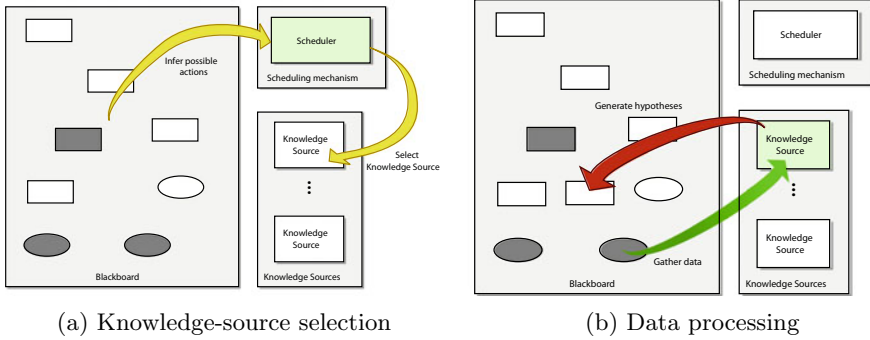


Fig. 3 Illustration of the blackboard-processing cycle. **a.** The scheduler selects the next knowledge source to be executed based on the current state of the blackboard. **b.** The selected knowledge source obtains data from the blackboard, processes it and writes the result back to the blackboard

2.3 Incorporation of Feedback Mechanisms

The basic blackboard processing framework introduced in Sect. 2.2 provides opportunities for possible extensions and adaptations. An important aspect that is of special relevance in dynamic problem domains like auditory scene understanding, is the incorporation of feedback mechanisms. Consider, as an example, an application in robot audition, where a dynamic agent equipped with acoustic sensors is able to freely explore its environment. A reasonable task for the robotic agent to perform in such an application is the localization and identification of all acoustic sources that are present in the immediate environment of the agent. Designing a blackboard system for this problem requires extensions of the model introduced in Sect. 2.2 as follows. Incorporate the ability of the robot to move about and provide capabilities to refine estimated source positions with identity estimates and vice versa. The blackboard system would then serve as an internal world model for the robotic agent, which, for instance, can be used for motion planning. These two exemplary extensions actually describe the following two different approaches to feedback loops that can be integrated into the blackboard system,

- *External* feedback loops, which directly affect the agent’s interaction with the environment, for example, the motion of a robotic agent or the rotation of a binaural dummy head. The initiation of these feedback loops will be controlled by the scheduler, implemented as additional actions that are available in the action space
- *Internal* feedback loops, describing incremental refinement processes within the blackboard itself. Improvements of source-localization estimates when additional information about source identities are available are examples that fit into this category—compare Ma et al. (2018), or, more generally, the priming of classifiers with additional information from the blackboard (Fig. 2).

3 Implementation of a Blackboard System with Hypothesis-Driven Feedback

To illustrate how an actual blackboard system can be implemented based on the framework presented in Sect. 2.2, an example from the domain of binaural localization based on the work proposed in Schymura et al. (2014) is given in current section. The blackboard system proposed by Schymura is broadly based on the HEARSAY-II Speech-Understanding System proposed by Erman et al. (1980). It comprises a blackboard data repository, a set of knowledge sources, and a scheduler, as introduced in the previous section. Additionally, the architecture is event-driven, that is, a change in the state of the blackboard, such as the arrival of new data, causes an event to be broadcast. A *blackboard monitor* is responsible for monitoring and handling these events. It maintains an *event register* that indicates which knowledge sources should respond to a certain event. The possible actions that can be performed, given the current state of the blackboard, are listed in an *agenda*. The scheduler is then responsible for ranking possible actions and selecting one of them to be performed. Completion of an action will most likely result in further changes in the state of the blackboard, leading to broadcast of new events.

The design of this blackboard system allows for a fusion of statistical and expert knowledge. The novel approach investigated here is the representation of knowledge by designing the blackboard as a set of interconnected graphical models, yielding a representation of the blackboard itself as a Bayesian network—see Pearl (1989). Computationally, this is realized by designing the blackboard to be a space for creating, assembling, and evaluating graphical models.

3.1 Motivation for a Graphical-Model-Based Architecture

Graphical models have attracted great interest in the fields of machine learning and of cognitive systems in general. They describe relationships between statistical variables in the form of simple graph structures. In these graphs, each node corresponds to a variable, and each edge indicates a dependency relationship between variables—see Bishop (2006). In this way, graphical models can be used to describe the dependencies between all variables that are of interest, effectively providing an interpretable world model.

Graphical models come in many different specific forms, such as hidden Markov models (HMM), Markov random fields, or dynamic state-space models that are suitable for creating precise descriptions of the constituent components of acoustic or audiovisual scenes. Efficient algorithms have been developed, which allow inferring latent variables of interest in an acoustic scene from observations taken available acoustic sensors. Hence, based on a graphical model of the audiovisual objects in an environment, the system will be able to find the best explanation of all available information, optimally fusing prior knowledge, such as linguistic or acoustic one,

with the currently available input from sensors. Taking graphical models as building blocks further allows one to

- Consecutively build models of the audiovisual environment from smaller, well-understood models of environmental objects—including state-of-the-art statistical models of auditory objects
- Understand sensory data as a composition of these source models and a model of the system’s own “perception”
- Understand the system’s interpretation of the audiovisual environment by virtue of the interpretability of each individual component and of their mutual connections.

Since the model is statistical in nature, the resulting interpretation of the environment will not only denote the type, number, location and—if applicable—the possible intention of all objects of interest, but also contain estimates of the corresponding uncertainties of all of these quantities. This will endow the system with the ability to judge the reliability of its own interpretation. This capability can ultimately be used to design and optimize active-listening and active-exploration processes to the end of ensuring that the most relevant variables are determined with sufficient reliability.

3.2 *Blackboard Architecture*

Figure 4 gives an overview on a general system architecture as proposed by Schymura et al. (2014) for solving a single-source speaker-localization task. This rather simple example has deliberately been chosen as a demonstration scenario with the purpose of serving as a proof-of-concept for the proposed architecture. The blackboard workspace is arranged into a hierarchy of four layers that are described in the following.

Acoustic-Cues Layer

The lowest layer, denoted as the *acoustic-cues layer*, contains observation vectors modeled as continuous, multivariate and observable random variables. The observations take the form of estimated interaural arrival-time differences (ITDs) and interaural level differences (ILDs) that can be added to the blackboard by the corresponding knowledge source, “Acoustic Cue KS”, that operates on this layer. The knowledge source takes the monaural left and right ear signals, acquired via a heard-and-torso simulator, here a KEMAR dummy head. These ear signals are processed by an auditory front-end, composed of an M -channel gammatone filterbank followed by an inner-hair-cell model—as proposed by Meddis (1986). This setup is used to subsequently estimate ITDs and ILDs independently at each time step, similar to the binaural-processing framework introduced by May et al. (2011). The resulting observation vector

$$\mathbf{y}_k = \left[\hat{\tau}_{k,1}, \dots, \hat{\tau}_{k,M}, \hat{\delta}_{k,1}, \dots, \hat{\delta}_{k,M} \right]^T \quad (2)$$

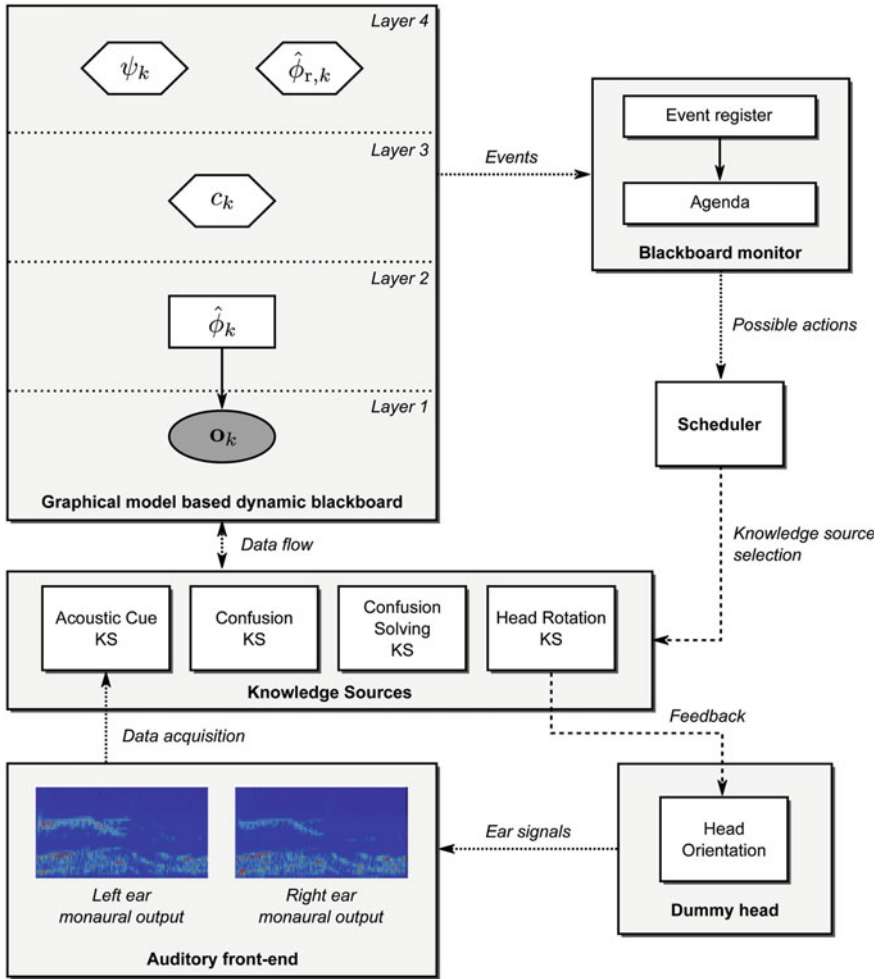


Fig. 4 Overview of the hierarchical blackboard architecture for active binaural localization introduced by Schymura et al. (2014). Data flow between the different components is represented by **dotted arrows**, whereas **dashed arrows** represent control commands. The different components on the blackboard are divided into continuous random variables (*ellipsoid nodes*), discrete random variables (*rectangular nodes*) and data segments (*hexagonal nodes*). The Gaussian mixture model that is used in layers 1 and 2 is illustrated by a *solid arrow* that represents the statistical relationship between the observation vectors o_k and the discrete locations $\hat{\phi}_k$

has the dimensionality $2M$, where $\hat{\tau}_{k,m}$ denotes the estimated ITD at the discrete time step, k , and the frequency channel, m with $m = 1, \dots, M$. $\hat{\delta}_{k,m}$, denotes the estimated ILD, respectively.

Localization-Hypothesis Layer

The central element of this layer of the blackboard architecture, which is referred to as the *location-hypothesis layer*, contains a discrete hidden random variable, $\hat{\phi}_k$, that represents hypotheses about the possible locations of a sound source in terms of azimuths. The variable $\hat{\phi}_k$ is statistically related to the corresponding observation vector described in (2). Both variables are linked via a Gaussian-mixture (GMM) model.

$$p(\mathbf{y}_k | \lambda_l) = \sum_{i=1}^N \pi_i p_i(\mathbf{y}_k | \lambda_l) \quad (3)$$

composed of N mixture components, with parameters $\lambda_l = \{\pi_i^{(l)}, \boldsymbol{\mu}_i^{(l)}, \boldsymbol{\Sigma}_i^{(l)}\}$. Each of the $l = 1, \dots, L$ Gaussian mixtures corresponds to a specific azimuth, ϕ_l . The mixture components in (3) are modeled as $2M$ -dimensional Gaussian distributions $p_i(\mathbf{y}_k | \lambda_l)$, with mean vectors $\boldsymbol{\mu}_i^{(l)}$, covariance matrices $\boldsymbol{\Sigma}_i^{(l)}$, and mixture weights $\pi_i^{(l)}$ satisfying $\sum_{i=1}^N \pi_i^{(l)} = 1 \forall l$. In this specific implementation, the number of Gaussian mixtures was limited to 72, yielding an angular resolution of 5° for the localization estimates. Whenever new observations are added to the blackboard, the Gaussian-mixture models are triggered to infer the posterior probability

$$p(\phi_l | \mathbf{y}_k) = \frac{p(\mathbf{y}_k | \lambda_l)}{\sum_{l'=1}^L p(\mathbf{y}_k | \lambda_{l'})} \quad (4)$$

of all possible azimuths. This process results in a categorical distribution over azimuth, $p(\hat{\phi}_k | \mathbf{y}_k)$, that is added to the blackboard. It should be noted that the utilized localization system is largely inspired by the work proposed by May et al. (2011), which was subsequently extended and improved—compare Ma et al. (2015) and May et al. (2015).

Confusion-Hypothesis Layer

To reduce localization errors caused by front-back confusion, a next layer is introduced in the blackboard architecture by the name of *confusion-hypothesis layer*. Front-back confusions are an inherent problem in binaural localization, which also occurs in human hearing—see Blauert (1997). Initial research in this direction was conducted by Wallach (1940), who found that humans exploit head rotations to resolve front-back confusions—compare Pastore et al. (2019). These findings have been adapted to machine-hearing systems in recent works of Ma et al. (2015) and Schymura et al. (2015). The blackboard system of Schymura et al. (2014), which is reviewed here, served as an initial step toward integrating the inherently active process of confusion solving into the blackboard architecture via hypothesis-driven

feedback. Confusion hypotheses are generated by the “Confusion KS”, which operates on the confusion hypothesis layer. This knowledge source examines whether location hypotheses of the second layer contain potential confusions. This examination is based on a threshold, $p_{\min} \in [0, 1]$, that defines a probability of which one of the posterior probabilities, $p(\hat{\phi}_l | \mathbf{y}_k)$, is considered as a location hypothesis. A confusion is identified when there are multiple location hypotheses within the same time step. When confusion is identified, a confusion hypothesis

$$c_k = \{\tilde{\phi}_{k,1}, \dots, \tilde{\phi}_{k,Q}\} \quad (5)$$

is created which includes all Q competing locations, $\tilde{\phi}_{k,j}$, $j = 1, \dots, Q$. If $Q = 1$.

Perceptual-hypotheses Layer

If no confusion is detected, a relative source-location hypothesis, $\hat{\phi}_{r,k}$, is created in this fourth layer of the blackboard, denoted as the *perceptual hypotheses layer*. The perceptual-hypotheses layer contains two variables, ψ_k and $\hat{\phi}_{r,k}$, corresponding to the current look-direction of the dummy head and the estimated relative source direction, respectively. As described before, if no front-back confusion was detected, the estimated relative source position is directly computed by the “Confusion KS” knowledge source from the posterior probabilities on the second layer. If there is a remaining confusion hypothesis on the third layer according to (5) and the head has not been rotated, the “Head Rotation KS” knowledge sources triggered. This halts the listening process and activates the feedback path that triggers a change of the current head orientation by 10° into the direction of the hypothesized sound source. After the rotation is completed, it indicates that the system is ready for the next time step and triggers the “Confusion Solving KS” knowledge source. This KS solves localization confusions by predicting the posterior probability of the source azimuth after a head rotation and by comparing it with new location hypotheses as have been gathered within the next time step. If a hypothesized source position reflects a “true” source location, then the predicted location distribution and the observed distribution after head rotation should overlap at the same location. If this is the case, the estimated position is considered a valid relative source location hypothesis, $\hat{\phi}_{r,k}$, which is then put onto the blackboard. The corresponding confusion hypotheses on the third layer are then discarded by the “Confusion Solving KS” knowledge source. If the predicted and observed distributions do not match, the hypothesized location is considered a *ghost*, and the system proceeds with the next frame to gather more data before repeating the process. An example of the confusion solving process is illustrated in Fig. 5.

The triggering of specific knowledge sources is attached to certain events that are stored in an event register, which is part of the blackboard monitor. As described before, events are generated if new data is available from the auditory front-end or if specific KS have performed certain actions on the blackboard. The blackboard monitor keeps track of the current state of the blackboard and generates an agenda which contains all actions that could be performed according to this state.

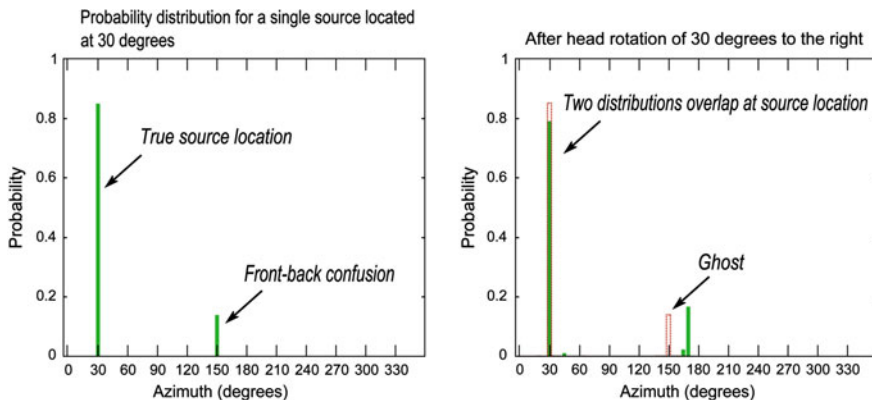


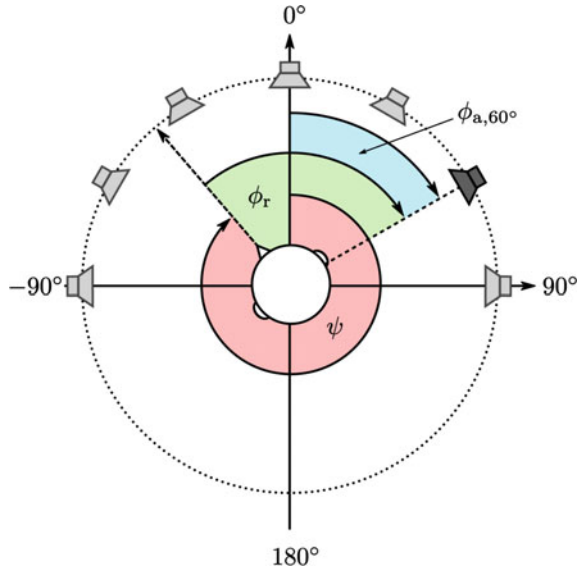
Fig. 5 Illustration of the front-back confusion solving mechanism—as proposed by Schymura et al. (2014). The **left panel** shows the probability distribution for different positions for a source located at an azimuth angle of 30° . There clearly exists a “ghost” at an angle of 150° . The **right panel** shows the predicted location distribution in **dotted lines** and the actual distribution after head rotation by 10° . The two distributions overlap at an azimuth angle of 30° which suggests a “true” source position

The agenda is then passed to the scheduler that decides which of the possible actions would be best suited given the current state of the blackboard and the task that should be accomplished. In the proposed system, a weight is attached to each knowledge source (KS) represented as a numerical value between 0 and 100. This weight corresponds to the importance of a specific KS for accomplishing the localization task. Given the agenda, the scheduler executes the action that is linked to the KS with the highest weight. This specific implementation of a blackboard system, utilizing a dedicated blackboard monitor with an attached agenda was adopted from Erman et al. (1980).

3.3 Experimental Evaluation

This section briefly reviews the results obtained in the experimental evaluation of the described system by Schymura et al. (2014). The presented blackboard system had the purpose to act a proof-of-concept. Therefore, a simple single-source localization scenario was chosen for the evaluation. An extension of the confusion-solving process to multi-source scenarios can be found in e.g., May et al. (2015). The position of the dummy head was assumed to be static, but changes in head orientation were possible. The target sound was a static speech source, located within the horizontal plane at an arbitrary azimuth between $[0^\circ, 360^\circ]$ with a 5° angular resolution. Since the localization task was not restricted to the frontal plane, the localization system also had to cope with potential front-back confusions.

Fig. 6 Illustration of the evaluation scenario utilized in Schymura et al. (2014). The azimuth, $\phi_{a,60^\circ}$, denotes an exemplary absolute target azimuth serving as the ground-truth. The relative angle between the look direction, ψ , of the dummy head and the target is denoted as ϕ_r . The dummy head can rotate within a range of $[-80^\circ, 80^\circ]$ in the frontal hemisphere



Seven target source positions were selected for the evaluation, namely, 270° , 300° , 330° , 0° , 30° , 60° , and 90° . Although the evaluated target positions were located only in the frontal sector of the horizontal plane, the localization system did not have this prior knowledge and assumed an azimuth range of $[0^\circ, 360^\circ]$ for potential target positions. The principal setup of the evaluation scenario is depicted in Fig. 6.

Localization performance was evaluated in two acoustic conditions. The first condition did not include any background noise, that is, only the clean target speech source was presented to the dummy head. This condition was evaluated to obtain an upper bound of performance, that is, where the best possible localization could be achieved. Additionally, the second condition included a diffuse noise field with a signal/noise ration (SNR) of 0 dB to evaluate the noise robustness of the proposed system. In both conditions, it was assumed that the listener and the sound source were located in free-field conditions. The simulation was generated using head-related transfer functions (HRTFs) of the dummy head, recorded at a distance of 3 m from the KEMAR. These data were taken from a database recorded by Wierstorf et al. (2011).

3.4 Experimental Setup

The target source was simulated using speech signals taken from the GRID corpus—introduced in Cooke et al. (2006). This database consists of short utterances spoken by 34 native English speakers (18 male speakers and 16 female ones). The training set utilized for estimating the Gaussian mixture included 340 randomly selected utter-

ances with 10 utterances per speaker. These were then spatialized to produce training data for each azimuthal position between $[0^\circ, 360^\circ]$ in steps of 5° . An additional set of 170 utterances with five utterances per speaker were selected as the evaluation set and were spatialized to simulate the seven target-source positions described above.

The diffuse noise field used in the second test condition was taken from the environmental sounds (“busy street”) from the IEEE AASP CASA Challenge Dataset—see Stowell et al. (2015). The noise was added to the binaural speech signals after spatialization at an SNR of 0 dB with respect to the averaged speech-signal power.

The peripheral processing of the auditory system was simulated by an auditory front-end proposed by May et al. (2011). The simulated ear signals were decomposed into 32 gammatone-filterbank channels. The center frequencies of the filterbank were equally distributed on the equivalent-rectangular bandwidth (ERB) scale between 80 Hz and 8 kHz. The channel output was then halfwave-rectified and used to extract channel-dependent binaural cues. A Hann window with a length of 20 ms was used for the analysis in each frame with an overlap between successive frames of 10 ms. The interaural arrival-time differences (ITDs) for each channel were estimated by choosing the maximum lag of a cross-correlation function within the range of $[-1, 1]$ ms. The interaural level differences (ILDs) in each channels were estimated by comparing the energy integrated across the window between the left and right ears within each channel and expressed in dB.

Two localization systems were evaluated, that is, a Gaussian-mixture-model-based localization baseline, which was unable to perform head rotations, and the proposed blackboard system incorporating the confusion-solving mechanism. Both systems used identical Gaussian-mixture models (GMMs) to model the azimuth-dependent distribution of the binaural feature space consisting of ITDs and ILDs. The GMM baseline selected the azimuth with the maximum posterior given a binaural feature observation as the target-source position, while the blackboard included top-down feedback for head rotation in order to resolve front-back ambiguities as described in Sect. 3.2. The GMMs were trained only on clean spatialized speech signals and no noise was included during training. No prior knowledge of source positions was used.

3.5 Results and Discussion

The localization performance of both systems was evaluated as utterance-level localization errors, which were computed by averaging the minimum angular differences between the reference target position and the estimated positions within each utterance. Table 1 shows the mean utterance-level localization errors based on the 170 test utterances for each evaluated target position.

Under clean conditions, both systems were able to localize the speech source at all the evaluated positions with very little error. A t -test with $p < 0.01$ showed that the performance of the blackboard system was significantly better than that of the GMM baseline. It should be noted that under clean conditions the GMM baseline

Table 1 Localization errors in degrees during the experimental evaluation in Schymura et al. (2014). The blackboard system incorporating hypothesis-driven feedback clearly outperforms the Gaussian-mixture (GMM) baseline in both clean and noisy conditions

Target azimuth	-90°	-60°	-30°	0°	30°	60°	90°	Avg.
GMM baseline (Clean)	0.07	0.15	0.27	0.23	0.14	0.60	0.16	0.23
Blackboard (Clean)	0.04	0.04	0.04	0.03	0.03	0.03	0.04	0.04
GMM baseline (Noisy)	8.32	17.63	29.94	0.73	6.87	11.41	7.94	11.84
Blackboard (Noisy)	0.13	0.80	1.43	0.04	0.41	0.42	0.25	0.50

was able to handle front-back ambiguities without head rotation. This is largely because the GMMs captured the azimuth-dependent patterns of binaural cues across all frequency channels. The subtle spectral difference between front and back was realized by the KEMAR HRTFs used in the simulation and thus implicitly modeled by the system.

When diffuse noise was present, the localization errors of the GMM baseline increased significantly across all target positions except for 0° azimuth. The performance was particularly bad for the GMM baseline at azimuth positions where the front-back confusion was strong, in particular, 30° and 60° at both sides. The performance of the blackboard system, however, was generally robust in the presence of the diffuse noise and was significantly better than the baseline. The top-down feedback that allowed head rotation helped the system resolve most ambiguities and the improvement over the baseline was consistent across all the target positions.

3.6 Summary

This chapter presented a review of the blackboard architecture proposed by Schymura et al. (2014), which extended the conventional blackboard design with graphical models and hypothesis-driven feedback mechanisms. It serves as a proof-of-concept for high-level auditory scene analysis frameworks, which, based on a graphical model representation, can iteratively develop an “understanding”—that is, an internal, interpretable description—of an auditory scene. While results were shown for a small toy example, namely the localization of a single acoustic source, the framework allows inference in a wide range of dynamic Bayesian networks, supporting many types of knowledge sources and inference strategies.

4 Current and Future Developments

In the previous section, a basic blackboard architecture designed for a simple auditory-scene-analysis task was reviewed. This was intended to outline the core components needed for implementing a functioning blackboard system for such

tasks, serving as a basis for further extensions. The model of Schymura et al. (2014) was extended within the TWO!EARS project¹ project, to create a flexible software architecture for active auditory scene understanding utilizing a mobile robotic agent. Notable extensions and improvements to the model presented in this chapter were the incorporation of sound type-detection (Trowitzsch et al. 2017), the refinement of source-localization estimates via spectral source models (Ma et al. 2018), as well as approaches for active exploration of acoustic scenes—see Bustamante et al. (2015, 2016a, b); Bustamante and Danès (2017); Schymura et al. (2017). A further notable implementation of the blackboard paradigm that was developed within the TWO!EARS project consortium is the virtual test environment for intelligent binaural models—compare Chap. 17, this volume. Such extensions can be implemented as additional knowledge sources which can then simply be added to the existing blackboard architecture. This allows for an adaptation of the blackboard system to specific tasks by adding suitable knowledge sources to the pool of already available ones. This is a clear advantage of blackboard systems over conventional machine-learning techniques, which are mostly suited to solve a specific task and cannot that easily be adapted to completely different tasks. However, recent developments in the field of machine-learning have also considered flexible architectures, which do not strictly follow the traditional blackboard paradigm but exhibit many similarities, especially regarding the external-control mechanism that is realized by the scheduler.

A prominent example are neural Turing machines proposed by Graves et al. (2014). Neural Turing machines are a specific implementation of neural networks that can access external memory. This memory access is controlled externally, in analogy to a conventional Turing machine. However, the proposed framework can be optimized via standard gradient descent, as the memory access is designed to be differentiable. A long short-term memory (LSTM) network is used as a control instance to focus the networks attention on specific areas in the external memory. Graves et al. (2014) report that the proposed system is able to infer simple algorithms, for instance, copying and sorting from examples. Even though neural Turing machines are fundamentally different from traditional blackboard architectures in most respects, the external control (“scheduler”) of the memory access (“blackboard”) are conceptually similar. This might also serve as a starting point for further research regarding blackboard architectures, where the scheduler might be represented by a neural network instead of using rule-based heuristics.

An extension to neural Turing machines are differentiable neural computers as proposed by Graves et al. (2016). In comparison to neural Turing machines, differentiable neural computers have greater flexibility in handling memory access and exploit a control mechanism that explicitly considers the order of memory read and write events. The underlying model is also fully differentiable, hence it can be trained via gradient descent. Differentiable neural computers can solve challenging tasks like processing complex data structures and symbolic-reasoning tasks—as were also considered during the early development of blackboard systems.

¹For detailed information, please refer to www.twoears.eu [last accessed August 24, 2019].

Recent developments in the field of machine learning, especially regarding progress concerning deep learning and neural networks provide interesting starting points for further investigations in the domain of blackboard architectures. Even though the classical models proposed in this field are not directly applicable to current tasks in machine learning, there are many possibilities for further research. Especially with the advent of emerging technologies in deep learning, probabilistic blackboard systems might serve as a starting point for designing systems that combine the best of both worlds.

Acknowledgements The authors thank two anonymous reviewers for thoughtful suggestions that helped to widen the scope of this chapter.

References

- Barot, V., M Henshaw, C Siemieniuch, and H. Dogan. 2013. Design of a web-based thesaurus for systems of systems engineering. In *2013 8th International Conference on System of Systems Engineering*, 7–12.
- Bishop, C.M. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin: Springer.
- Blauert, J. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press.
- Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge: MIT Press.
- Bustamante, G., and P. Danès. 2017. Multi-step-ahead information-based feedback control for active binaural localization. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6729–6734.
- Bustamante, G., P. Danès, T. Forgue, and A. Podlubne. 2016. A one-step-ahead information-based feedback control for binaural active localization. In *2016 24th European Signal Processing Conference (EUSIPCO)*, 1013–1017.
- Bustamante, G., P. Danès, T. Forgue, and A. Podlubne. 2016. Towards information-based feedback control for binaural active localization. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6325–6329.
- Bustamante, G., A. Portello and P. Danès. 2015. A three-stage framework to active source localization from a binaural head. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5620–5624.
- Cooke, M., J. Barker, S. Cunningham, and X. Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120(5): 2421–2424.
- Cooke, M., G.J. Brown, M. Crawford, and P. Green. 1993. Computational auditory scene analysis: listening to several things at once. *Endeavour* 17(4): 186–190.
- Corkill, D.D. 1991. Blackboard systems. *AI Expert* 6: 40–47.
- Dirand, J., and V. Chevrier. 1995. A blackboard-based expert system for engineering design of electrical transformers. In *Proceedings 1995 Canadian Conference on Electrical and Computer Engineering*, vol. 2, 784–787.
- Ellis, D.P.W. 1996. Prediction-driven computational auditory scene analysis. Ph.D. thesis, Massachusetts Institute of Technology.
- Erman, L.D., F. Hayes-Roth, V.R. Lesser, and D.R. Reddy. 1980. The Hearsay-II speech understanding system: integrating knowledge to resolve uncertainty. *Computing Surveys* 12(2): 213–253.

- Fox, C., M. Evans, M. Pearson, and T. Prescott. 2012. Towards hierarchical blackboard mapping on a whiskered robot. *Robotics and Autonomous Systems* 60(11): 1356–1366 towards Autonomous Robotic Systems 2011.
- Godsill, S., and M. Davy. 2002. Bayesian harmonic models for musical pitch estimation and analysis. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, II–1769–II–1772.
- Godsmark, D., and G.J. Brown. 1999. A blackboard architecture for computational auditory scene analysis. *Speech Communication* 27(3–4): 351–366.
- Graves, A., G. Wayne, and I. Danihelka. 2014. Neural turing machines. *CoRR* <https://arxiv.org/abs/1410.5401> (last accessed Dec. 2019).
- Graves, A., G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwinska, S.G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou, A.P. Badia, K.M. Hermann, Y. Zwols, G. Ostrovski, A. Cain, H. King, C. Summerfield, P. Blunsom, K. Kavukcuoglu, and D. Hassabis. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature* 538 (7626): 471–476.
- Hayes-Roth, B. 1985. A blackboard architecture for control. *Artificial Intelligence* 26(3): 251–321.
- Hewett, M., and R. Hewett. 1993. A language and architecture for efficient blackboard systems In *Proceedings of 9th IEEE Conference on Artificial Intelligence for Applications*, 34–40.
- Hong, X., C. Xinrong, L. Qun, and Z. Jianpei. 1997. Knowledge-based diagnostic system of turbine with faults using the blackboard model. In *1997 IEEE International Conference on Intelligent Processing Systems (Cat. No.97TH8335)*, vol. 2, 1516–1519.
- Hou, P.K., X.Z. Shi, and L.J. Lin. 2000. Generic blackboard based architecture for data fusion. In *2000 26th Annual Conference of the IEEE Industrial Electronics Society. IECON 2000. 2000 IEEE International Conference on Industrial Electronics, Control and Instrumentation. 21st Century Technologies*, vol. 2, 864–869.
- Lesser, V.R., S.H. Nawab, F.I. Klassner. 1995. IPUS: An architecture for the integrated processing and understanding of signals. *Artificial Intelligence* 77: 129–171.
- Ma, N., J.A. Gonzalez, and G.J. Brown. 2018. Robust binaural localization of a target sound source by combining spectral source models and deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26(11): 2122–2131.
- Ma, N., T. May, H. Wierstorf, and G.J. Brown. 2015. A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2699–2703.
- May, T., N. Ma, and G.J. Brown. 2015. Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2679–2683.
- May, T., S. van de Par, and A. Kohlrausch. 2011. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech, and Language Processing* 19(1): 1–13.
- McManus, J.W., and W.L. Bynum. 1996. Design and analysis techniques for concurrent blackboard systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 26(6): 669–680.
- Meddis, R. 1986. Simulation of mechanical to neural transduction in the auditory receptor. *The Journal of the Acoustical Society of America* 79(3): 702–711.
- Nii, H.P. 1986. Blackboard systems. Technical Report. Stanford, CA, USA: Stanford University. <http://i.stanford.edu/pub/cstr/reports/cs/tr/86/1123/CS-TR-86-1123.pdf> (last accessed Dec. 25, 2019).
- Pang, G.K.H. 1988. A blackboard control architecture for real-time control. In *1988 American Control Conference*, 221–226.
- Pastore, M., Y. Zhou, and W.A. Yost. 2019. Cross-modal and cognitive processes in sound localization. In *The Technology of Binaural Understanding*, eds S. Blauert, J. and J. Braasch, 315–350. Cham, Switzerland: Springer.

- Pearl, J. 1989. Morgan Kaufmann series in representation and reasoning *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Burlington: Morgan Kaufmann.
- Rong, Z. 2014. Information system of emergency materials management based on active blackboard structure design. In *2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications*, 570–573.
- Schymura, C., J.D.R. Grajales, and D. Kolossa. 2017. Monte carlo exploration for active binaural localization. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 491–495.
- Schymura, C., N. Ma, G.J. Brown, T. Walther, and D. Kolossa. 2014. Binaural Sound Source Localisation using a Bayesian-network-based Blackboard System and Hypothesis-driven Feedback. In *Proceedings of Forum Acusticum*, Kraków, Poland.
- Schymura, C., F. Winter, D. Kolossa, and S. Spors. 2015. Binaural sound source localisation and tracking using a dynamic spherical head model. In *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 165–169.
- Song, J.J., K.H. Park, and S.W. Park 1991. A blackboard-based expert system for control structure synthesis of chemical processing plants. *Journal of Intelligent Manufacturing* 2(6): 379–385.
- Stowell, D., D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley. 2015. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia* 17(10): 1733–1746.
- Sutton, C., Morrison, C., Cohen, P., Moody, J., Adibi, J. (2004). A Bayesian Blackboard for Information Fusion. In *Proc. 7th International Conference on Information Fusion*, <http://fusion2004.foi.se/papers/IF04-1111.pdf> [last access: Dec. 27, 2020].
- Sutton, R.S., and A.G. Barto. 1998. *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press.
- Trowitzsch, I., J. Mohr, Y. Kashef, and K. Obermayer. 2017. Robust detection of environmental sounds in binaural auditory scenes. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25(6): 1344–1356.
- Wallach, H. 1940. The role of head movement and vestibular and visual cues in sound localization.
- Wang, D., and G.J. Brown. 2006. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.
- Wang, Y., X. Li, and J. Tian, 2011. Blackboard mechanism based genetic algorithm for dynamic deployment of mobile sensor networks. In *Proceedings of 2011 International Conference on Electronic Mechanical Engineering and Information Technology*, vol. 6, 2796–2799.
- Weiss, M., and F. Stetter. 1992. A hierarchical blackboard architecture for distributed AI systems. In *Proceedings Fourth International Conference on Software Engineering and Knowledge Engineering*, 349–355.
- Wierstorf, H., M. Geier, and S. Spors. 2011. A free database of head related impulse response measurements in the horizontal plane with multiple distances. In *Proceedings of the 130th Audio Engineering Society Convention*.
- Zhu, T., G. Liu, and L.M. Jia. 2010. A cooperative making multi-agent model on railway daily dispatching plan based on blackboard. In *2010 International Conference on Digital Manufacturing Automation*, vol. 1, 18–21.

Configuring and Understanding Aural-Space

Formation of Three-Dimensional Auditory Space



Piotr Majdak, Robert Baumgartner and Claudia Jenny

Abstract Human listeners need to permanently interact with their three-dimensional (3D) environment. To this end, spatial hearing requires efficient perceptual mechanisms to form a sufficiently accurate 3D auditory space. This chapter discusses the formation of the 3D auditory space from various perspectives. The aim is to show links between cognition, acoustics, neurophysiology, and psychophysics. The first part presents recent cognitive concepts for creating internal models of the complex auditory environment. Second, the acoustic signals available at the ears are described and the spatial information they convey is discussed. Third, neural substrates forming the 3D auditory space in the brain are explored. Finally, the chapter elaborates on psychophysical spatial tasks and percepts that are only possible because of the formation of the auditory space.

1 Introduction

A loud roar—do you turn around? Decide quickly: fight or flight? This archaic situation was typical for human ancestors still living in their natural habitats. Nevertheless, it still applies to the modern human: You hear the expression “Hi!” and turn around—a good friend has just recognized you on the street, but before you change your walking direction, a car honk lets you look back—you have just missed that car crossing your path. Situations like this one make it obvious: In the jungle and on the street, humans need a good understanding of the 3D world through auditory perception. To this end, the human brain creates maps of the environment. Towards this goal, the auditory system helps in answering the question: “What is where?”

This chapter reviews recent advances in understanding the formation and the usage of the auditory space by human listeners—from the cognitive, acoustic, neurophysiological, and psychophysical perspectives. Section 2 describes the problem and elaborates on the potential solutions given by researchers from cognitive psychology.

P. Majdak (✉) · R. Baumgartner · C. Jenny
Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
e-mail: piotr.majdak@oeaw.ac.at

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_5

The perceptual outcome depends on the quality with which the acoustic spatial information is conveyed, thus, Sect. 3 highlights how this information is encoded within binaural signals. The solution to the problem is reflected in the neural processes of the auditory system, and thus, in Sect. 4, the neural processing of the acoustic signals is described briefly while focusing on the extraction of spatial information. Finally, the results of these processes manifest in the variety of spatially oriented tasks human listeners can complete. Hence, Sect. 5 reviews various psychophysical spatial tasks demonstrating the human ability to utilize the understanding of the 3D auditory world.

2 Cognition: Representing the World

From the cognitive perspective, understanding the world involves the formation of a mental representation of the environment. Understanding here means answering the question: “What is where?” However, sound is ephemeral, namely, it is happening and short-lived; it is an effect of events, providing information on what happens right now, instead of a long-lasting description of objects’ properties. Thus, the information carried by the sound not only needs to be stored for its processing, but it also requires consideration of various time scales. On a short time scale, a creak might mean someone stepping on the floor. Many similar creaks, however, would rather indicate someone opening a door. The ephemeral property of sound requires the auditory system to address the question: “What is happening?” But even answering this question can only help in understanding a *static* environment. The world is *dynamic* and full of interaction. In order to decide on its actions, the auditory system has to provide a basis for the prediction of “what will happen next?”

2.1 *The Ill-Posed Problem*

Auditory scenes consist of objects producing sounds, as illustrated in Fig. 1a. Perception is a process of transforming sensory information to higher levels of representation, as a means to represent these objects and their properties mentally. Here, an *auditory object* can be thought of as a perceptual construct linking a sound with a corresponding source (Griffiths and Warren 2004). Sitting in a park, hearing a honk, a word, and a chirp would let people identify auditory objects representing a car, a human, and a bird. Auditory objects have a spatial position as a property among other non-spatial properties. Now, when the human starts to speak, and the bird starts to sing, these two objects become sources of *acoustic* streams. Their mixture arrives at a listener’s ears, and the job of the auditory system is to separate them into two *auditory* streams, which can be defined as a series of coherent events grouped and attributed to a single auditory object. Hence, the formation of the auditory space

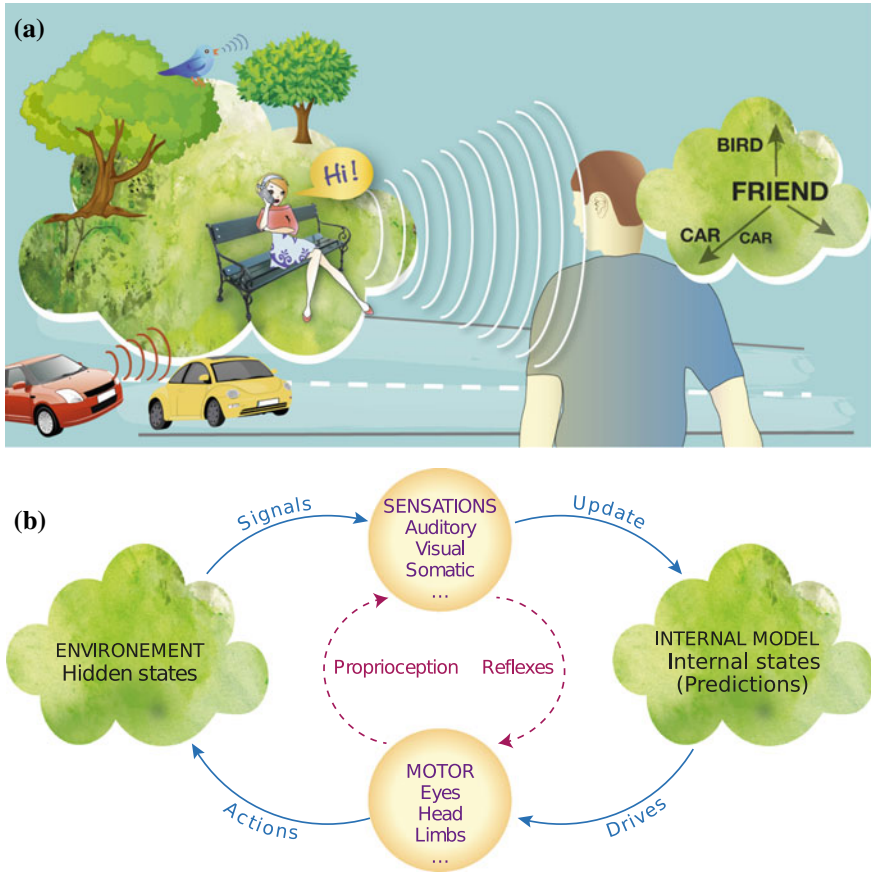


Fig. 1 **a** Perception as a mental representation of the environment based on sensation. The objects (left) produce sounds that are perceived by a listener (center) and are represented in an internal model (right). **b** Active inference. Sensations (top) of signals from the environment (left) are used to update the internal model predictions (right) on the hidden states of the environment. The model predictions drive actions (bottom) which allow the listener to interact with the environment

depends on the capabilities of the listener to form auditory objects and estimate their spatial properties.

Most of the sound sources are usually located outside of the body, so the formation of the auditory space can be seen as a perceptual task aiming at the reconstruction of the external (or distal) state of affairs (Epstein and Rogers 1995). Unfortunately, this is an ill-posed problem emerging from the fact that the information available at the ears is insufficient to reconstruct a unique state exactly because a given binaural signal, arriving at the ear drums, can originate from an infinite number of sound-producing events, all of which have produced exactly that signal. As a consequence, this ill-posed problem results in ambiguity when creating auditory objects.

A famous example showing the difficulty of mapping a sound to the source producing that sound is the estimation of a 2D surface based on the sound wave only. This problem has been posed as: “Can we hear the shape of a drum?”, and it has been formally shown that for many shapes, one cannot completely differentiate the shape of the drum because a unique reconstruction of the plane geometry from the waveform is impossible (Gordon et al. 1992). Fortunately, the acoustic environment is full of redundancy and recent developments in mathematical tools show the variety of tricks the auditory system may potentially use when it comes to utilizing acoustic redundancy to better solve inverse problems. For example, with a known monophonic signal, the shape of any convex room can be estimated just from the delays between the early reflections in the room impulse response (Moore et al. 2013). With four acoustic sensors, the room shape can be estimated by relying on the delay of just the *first* reflections (Dokmanić et al. 2013), showing the benefit of increased redundancy by using multiple sensors. Thus, it is not surprising that *binaural* as compared to monaural hearing allows the auditory system to better retrieve the spatial properties of the environment, helps to orient itself within the environment, and even improves speech perception by providing interaural information—see Bronkhorst (2015) and Clapp and Seeber (2020), this volume.

2.2 *Predictive Coding and Active Inference*

Despite the redundancy in the binaural signal, solving the ill-posed problem requires certain assumptions in order to reduce the infinite number of potential solutions (e.g., Friston 2012). Generally, these assumptions are driven by the goal of efficient interaction with the environment sampled by the sensors. At the end of the cognitive process, the solution needs to provide a basis for decisions that trigger the execution of appropriate actions. If the vast amount of sensory information were processed from scratch each time an action was required, most of the actions would happen too late. Faster processing can be achieved with *predictive coding* by introducing an *internal model* (Francis and Wonham 1976) that predicts the external state of affairs and is continuously adapted based on incoming sensory data, as shown in Fig. 1b—for a detailed review on predictive coding, see Aitchison and Lengyel (2017). Such models have been introduced in motor-control theory and robotics to describe reaching movements, to plan movement trajectories, and to model imagery—Grush (2004) reviews and discusses this topic extensively. In cognition, the term *perceptual inference* has been coined (Hinton and Ghahramani 1997).

In predictive coding, the more realistic the model predictions are, the more efficiently actions can be performed. In other words, the objective of the internal model is to minimize surprise that then requires only tiny corrections to be applied to the performed actions. In that sense, the process of cognition can be considered as forming a generator creating hypotheses that are tested against the pre-processed sensory information (Gregory 1980). The *free-energy principle* has been proposed to explain how the cognitive system can efficiently create a model predicting the environment

while restricting itself to a limited number of states (Friston et al. 2006). Free energy, a concept with a long tradition in thermodynamics (Helmholtz 1954), is the difference between the internal energy of a system and the energy required to describe the actual states of that system. From a cognitive perspective, free energy scores surprise and uncertainty about a belief. Hence, less free energy stands for a more certain and efficient system description. Based on the sensory information and model's beliefs about the environment, a cognitive model creates internal states that minimize surprise and thus the free energy. The free energy acts as a prediction error and is minimized by choosing the prediction that is most plausible and as such most efficiently drives the motor system. The internal states of the model are then updated based on the new sensory information about the hidden states of the environment. This process has been termed *active inference*—see Friston et al. (2016) for more details on this topic.

In active inference, a model's beliefs represent rules describing plausible environments. They can be learned throughout the development of an individual (Bhatt and Quinn 2011). The learned rules limit the potential solutions to plausible scenarios, having several implications. First, they sometimes fail, yielding unrealistic representations. Illusions, i.e., distortions of the perceived physical reality, are great examples of the consequences of plausible but wrong assumptions while solving the ill-posed problem (e.g., Carbon 2014). Understanding their origin can help to uncover the underlying processes in auditory perception. Second, these limitations reduce the vast amount of sensory information to a smaller number of informational units along the ascending pathways of processing. In this case, a small and discrete number of auditory objects with a finite number of properties are created from a continuous binaural signal. Thereby, the frame of reference is transformed from the head-centered binaural information to world-centered information about the environment (Schechtman et al. 2012).

Interestingly, this whole process can be seen as a nonlinear extension of *compressed sensing*, a signal processing technique for efficiently acquiring and reconstructing a signal by finding solutions to underdetermined linear systems (Donoho 2006). In compressed sensing, the constraint of sparsity is chosen in order to find a solution to the underdetermined system. Compressed sensing is widely used in signal processing, but it requires a *linear relation* between the observation and solution. Active inference, instead uses variational *Bayesian statistics*, to infer the unobserved variables based on an analytical approximation of their posterior probability (beliefs)—compare Friston et al. (2016).

2.3 Auditory Scene Analysis

In the end, it is all about reducing the amount of sensory information. A widely accepted concept describing the reduction of auditory information to discrete informational units is termed *auditory scene analysis* (ASA, Bregman 1990, and van de Par et al. 2020, this volume). ASA assumes that the auditory system partitions the

acoustic signal into auditory streams that each refer to an auditory object. While good separability between fore- and background streams was originally believed to constitute the goal (Bregman 1990), more recent models consider their predictability as the main motivation for grouping (Winkler et al. 2009). Grouping mechanisms seem to rely on auditory features like onset, pitch, spectrum, and interaural disparity and can act simultaneously and sequentially. Simultaneous grouping assumes that features are integrated to a foreground property—for example, harmonics coming from the same instrument are integrated to a single pitch—and features deviating from the expected and learned patterns are segregated and form a background. Sequential grouping integrates and segregates auditory objects and streams, depending on their temporal properties. This grouping effect can even override the result of simultaneous grouping. For example, two sounds having different interaural disparities, when presented simultaneously, can be grouped into a single auditory stream and can be perceived as a single auditory object appearing at a single spatial location. But, the same sounds, embedded in an acoustic stream having the interaural disparity corresponding to one of these sounds, can be perceived as *two* auditory objects appearing at *two* distinct spatial locations (Best et al. 2007).

Bregman's ASA concept of the "old-plus-new" strategy for competitive processes also determines how an auditory stream is formed (Bregman 1990). These processes were originally derived with respect to the laws of the *Gestalt theory*, which provides a description of the ability to acquire plausible perceptions from the sensory input (Koffka 1935). The main assumption of the Gestalt theory is that perception is based on grouping the sensory information to perceptual units according to the laws of proximity, similarity, closure, common fate, continuity, good form, and experience. Even though the Gestalt theory has difficulties in providing insights into the neural processes leading to perception (Schultz and Schultz 2015), the laws of the Gestalt theory helped in constraining the ambiguity resulting from the ill-posed problem of perception. These neural processes are based on heuristics acquired through learning and experience (Shinn-Cunningham 2008). Further, these processes are *top-down* mechanisms and can be modulated by other modalities like vision (yielding ventriloquism or self-motion; Kondo et al. 2012) or by attention (Hill and Miller 2010; Deng et al. 2019).

Depending on the relevance of top-down modulations, neural processes can be distinguished as being reflexive or reflective—see Blauert and Brown (2020), this volume, on their definitions and context. *Reflexive processes* result in speedy reactions (with latencies below 100 ms), which can usually not be suppressed (Curtis and D'Esposito 2003). They involve the startle reflex or orienting reflex (Sokolov 2001), and do not require, but can be modulated by attention. They can be used to trigger movements toward auditory sources or intensify the processing of cues that signal approaching objects (Baumgartner et al. 2017). Thus, they are vital in protecting humans from hazardous events. In contrast, *reflective processes* have longer latencies and require top-down attention, namely, a controlled bias in the preference for and processing of the information streams—see Knudsen (2007) for a review on this topic. Attention is usually thought to be a single, unidirectional top-down process representing task-specific goals and expectations (Awh et al. 2012). However, it

can also be modulated through a bottom-up mechanism by various components of a stimulus (Arnal et al. 2015). Reflective or attentional processes require working memory to develop and test hypotheses based on the salience in a stimulus (Carlile and Corkhill 2015). A similar distinction between reflexive and reflective processes has also been proposed for speech category learning (Chandrasekaran et al. 2014) and to describe social behavior (Strack and Deutsch 2004).

While the reflexive processes in auditory processing have been widely investigated, the actual reflective processes have not been completely understood yet. Thus, it is not surprising that mechanisms underlying ASA have been widely discussed from different perspectives (e.g., Bizley and Cohen 2013; Szabó et al. 2016; Micheyl et al. 2007; Nelken et al. 2014; Snyder and Elhilali 2017) and they build a foundation for what is known as *computational auditory scene analysis* (CASA, e.g., Wang and Brown 2006).

In summary, the free-energy principle provides a solid statistical framework for deriving the external state of affairs, and ASA provides a valid conceptual framework for the cognitive processing of auditory information. The particular result in terms of a realistic representation of the world depends on the quality with which the binaural signal conveys the spatial information about the auditory objects. Thus, the following section describes the acoustic spatial information encoded in binaural signals.

3 Acoustics: Formation of Binaural Signals

The sounds arriving at the ear drums are acoustically filtered versions of the sounds produced in the environment. The filtering results from the interaction of the sound field with the reverberant space and the listener's body parts such as the head, torso, pinnae, and ear canals. This section focuses on the acoustic effects of that filtering.

In acoustics, sound fields are commonly described in Cartesian or spherical coordinates. Aiming at disentangling the different cues contributing to the auditory localization of a sound source, a special form of spherical coordinates was established, the so-called *interaural-polar coordinate system*, as shown in Fig. 2a. In this system, the coordinate origin is located midway between a listener's ears. The poles of the spherical system are aligned with the interaural axis connecting the two ears. The lateral angle α describes the lateral position of a source. It ranges from -90° (right side) to $+90^\circ$ (left side) and selects a sagittal plane, that is, a plane parallel to the median plane. The lateral angle also corresponds to the azimuth angle of the frontal half of the the horizontal plane. The polar angle β describes the position of the source along the sagittal plane and ranges from -90° (bottom) via 0° (eye-level, front), 90° (top), and 180° (eye-level, back) to 270° (bottom again). Together with the distance r , lateral and polar angles are used throughout this chapter to describe the spatial position of sound sources.

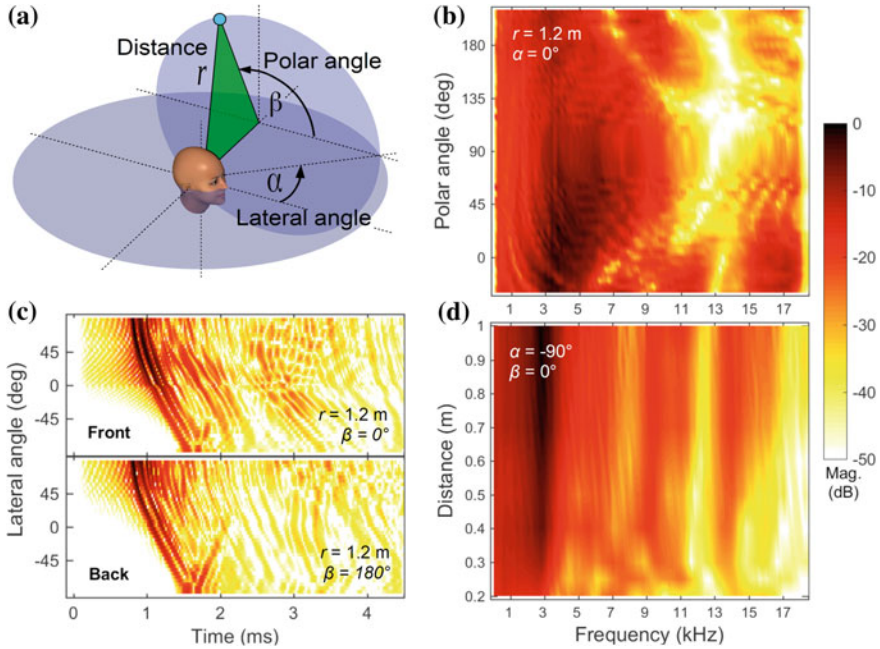


Fig. 2 **a** Interaural-polar coordinate system with the lateral angle α , polar angle β , and distance r ; **b** Magnitude spectra of far-field left-ear HRTFs along the median plane as a function of the polar angle; **c** Energy-time curves of far-field left-ear HRIRs along the lateral angle for the front and back of the horizontal plane; **d** Magnitude spectra of near-field left-ear HRTFs for the rightmost direction of the horizontal plane as a function of distance, compensated for the distance-related broadband magnitude decrease of a point source approximating the sound source

3.1 Monaural Features

For a sound coming from a particular direction, its filtering can be captured by the binaural pair of *head-related transfer functions* (HRTFs). While the filtering of the sound happening in the ear canal does not depend on the incidence angle of the sound within the normal hearing range, the filtering by the head, torso, and pinnae creates direction-dependent changes to the received sound. Spectral changes are especially apparent for a sound moving along the median plane of the listener, as shown in Fig. 2b. There is a clear relationship between changes in polar angle and the resulting spectrum; notches and peaks arise as a consequence of cancellation and amplification, caused by various body parts. The reflections of the torso create spatial frequency modulations up to 3 kHz (Algazi et al. 2001a). The head shadows frequencies above 1 kHz and above 6 kHz, the effect of the pinna is most prominent (Blauert 1997) and reflections at the pinna create distinct peaks and notches. For example, the directionality of the pinna towards the front causes high-frequency attenuation for sounds coming from behind the listener. This manifests in increased

energy above 8 kHz for the frontal sound positions. Further, an increase in sound elevation changes the varying delay between the direct sound and its reflection along the pinna concha. This manifests in an upward shift of the spectral notches usually found between 6 and 10 kHz.

In general, the individual shape of the pinna is the reason for a considerable variation in the HRTFs among listeners, as the pinna's geometry largely varies among the human population (Algazi et al. 2001a). While HRTFs are similar across listeners at frequencies up to 6 kHz, differences as large as 20 dB have been found at higher frequencies (Møller et al. 1995). Listener-specific HRTFs can be acquired by applying system identification approaches on acoustical measurements—see Majdak et al. (2007) for a review on this topic. The acoustic measurement is a resource-demanding procedure when it includes many positions in 3D space. It takes tens of minutes, even when sophisticated measurement methods are applied. HRTFs can also be calculated based on a geometric representation of the listener (Katz 2001; Kreuzer et al. 2009); however, the demands on geometric accuracy and computational power are high (Ziegelwanger et al. 2015). Recent developments in the acquisition of the 3D geometry from photographs using photogrammetric reconstruction (Reichinger et al. 2013) and numeric algorithms (Ziegelwanger et al. 2016) seem promising in easing the acquisition of listener-specific HRTFs in the future. While HRTFs have been measured for a long time for research purposes, their exchangeability was limited because of missing standards for their representation. The *spatially oriented format for acoustics* (SOFA) was created (Majdak et al. 2013a) as a standard of the Audio Engineering Society (AES69-2015 2015) in order to simplify their exchangeability and promote their usage in consumer applications.

HRTFs can also be analyzed in the time domain by applying the inverse Fourier transformation on each HRTF yielding *head-related impulse responses* (HRIRs), as shown in Fig. 2c. While both terms, HRIRs and HRTFs, can be used to describe the directional filtering interchangeably, the particular choice depends on the focus on time and frequency domain, respectively. HRIRs usually decay within the first 4 ms and show the direction-dependent delay between the sound source and the ear. The temporal position of the first onset in an HRIR can be considered as the broadband time-of-arrival (TOA). Based on the approximation of the listener's head as a sphere (Algazi et al. 2001b), the TOA can be described as a spatially-continuous function requiring only a few parameters, such as the direction-independent TOA, the head radius, and the ear position (Ziegelwanger and Majdak 2014). Even though HRTFs show a nonlinear spectral phase depending on the sound direction, the HRTF phase spectrum has been represented by a combination of the minimum phase derived from the HRTF amplitude and the linear phase corresponding to the TOA (Kulkarni et al. 1999).

HRTFs also vary with distance, especially in the near field, as shown in Fig. 2d. This is due to the contribution of the head shadow and changes of the pinna-reflection paths (Brungart and Rabinowitz 1999). The nearer the source, the less diffraction around the head occurs at lower frequencies, and the less intense are the reflection patterns of the pinna. Low-frequency attenuation of up to 20 dB is a prominent spatial feature encoding distance for near sounds.

The sound source itself might be a source of spatial cues because the HRTF filtering is commutative and the auditory system has no chance to distinguish whether the evaluated spectrum originates from the sound source or is an effect of filtering by an HRTF. For example, a narrow band sound can have a spectral similarity to a signal of a broadband source filtered by an HRTF, and a frequency sweep of notched noise can be similar to a signal created by a source moving in elevation. Both observations raised the hypothesis of directional frequency bands (Blauert 1969). Alternatively, sounds with spectral ripples may overlap with the monaural spectral features of HRTFs and interfere with the derivation of directional information from the binaural signal (Macpherson and Middlebrooks 2003).

3.2 Interaural Cues

Having two ears allows listeners to probe the sound field at two different spatial positions. Thus, listeners have access not only to monaural features but also to the combination of the left- and right-ear features, the so-called interaural cues. The two major interaural cues are the *interaural time differences* (ITDs) and *interaural level differences* (ILDs). Their importance for sound localization was recognized very early (Strutt alias Lord Rayleigh 1876). Later, the general dissimilarity of the signals between the two ears expressed as binaural incoherence and spectral ILDs, has been found to be important—see Blauert (1997) for more details on this topic.

3.2.1 Interaural Signal Similarity

The similarity between the signals of the listener's two ears can be described by the *interaural cross-correlation function* r_{LR} of the two signals x_L and x_R as a function of the interaural lag τ (Goupell and Hartmann 2006):

$$r_{LR}(\tau) = \frac{\int_{-T}^T x_L(t)x_R(t + \tau) dt}{\sqrt{\int_{-T}^T x_L^2(t_1) dt_1 \int_{-T}^T x_R^2(t_2) dt_2}}. \quad (1)$$

Also, other terms have been used to describe binaural similarity, for example, *interaural coherence* function (the Fourier transform of the interaural cross-correlation function), *binaural incoherence*, or interaural decorrelation. The interaural cross-correlation function $r_{LR}(\tau)$ can be computed with different integration time windows T ; often, a time window of about 1 ms is used to consider the naturally occurring ITDs. The function $r_{LR}(\tau)$ varies between -1 and 1 and typically has a single peak. The lag of that peak corresponds to the broadband ITD and is mostly determined by the lateral position of a sound source. The height of that peak, namely, $\max(r_{LR})$, is usually known as the *interaural cross-correlation coefficient* (IACC) and demonstrates the best interaural similarity of the binaural signal, as shown in Fig. 3a. The

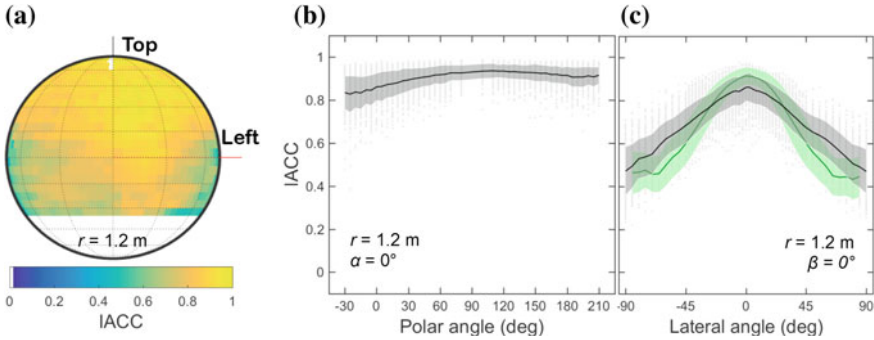


Fig. 3 **a** Interaural cross-correlation coefficients (IACCs) of a single listener’s HRTFs for various directions when looking at the listener from the front. **b** IACCs along the median plane calculated for a listener population. **c** IACCs along the horizontal front (gray) and rear (green) half-planes calculated for a listener population. The population consisted of 97 listeners from the ARI database (Majdak et al. 2010). The dots show the individual IACCs, the line and gray area shows the total average and ± 1 standard deviation, respectively, across the population

terms coherence and IACC are often used interchangeably, however, note the difference to the term *correlation coefficient*, that refers to $r_{LR}(0)$. Even in the median plane, free-field IACCs are typically below one, as shown in Fig. 3b, because pinnae are not perfectly symmetrical which yields tiny interaural differences even in the median plane. In the horizontal plane, the IACC decreases with increasing lateral angle of the sound source, with typical IACCs around 0.4 for the most-lateral directions, as shown in Fig. 3c. Note that this is a broadband consideration of the IACC and the interaural dissimilarities may be different and contribute differently across frequencies.

3.2.2 Interaural Time Differences

When dealing with ITDs, people usually refer to the broadband ITDs. However, theoretical considerations show that low-frequency ITDs are 50% larger than those at higher frequencies (Kuhn 1977) with a transition frequency around 1.5 kHz. Thus, given the frequency dependence and thus the limited interaural coherence, a broadband ITD is only an approximation of spectral delays appearing between the two ears. Consequently, various methods have been proposed for the estimation of ITDs. Figure 4a shows measured ITDs as a function of the lateral angle obtained from the HRTFs of a listener by using various methods. In the time domain, methods either evaluate the ITD between the first onsets (MAX in Fig. 4a), centroids (CTD), or the lag of the coherence function peak of an HRIR compared to its minimum-phase version (MCM). In the frequency domain, the ITD can be calculated from the spectral average of the interaural group delay (AGD). The best estimation method depends on the application. While ITDs between the onsets (30 dB below the peak) of low-pass

filtered HRIRs were found to best correspond with results from a psychoacoustic method of adjustment (Andreopoulou and Katz 2017), the MCM method was shown to provide highest geometrical consistency (Ziegelwanger and Majdak 2014). From the geometrical perspective, there is a long history of various ITD models based on representations of the head as a circle, sphere, ellipsoid, and polynomial function—see Xie (2013) for a review on this topic. As an example, Fig. 4a also shows the ITDs reconstructed by the spatially continuous 3D model of TOAs using the MAX method (Ziegelwanger and Majdak 2014).

The maximum ITD depends on the listener's head diameter and the calculation method and has a population average of around $850 \mu\text{s}$ (Algazi et al. 2001a). ITDs in that range imply that sounds with frequencies below 1.2 kHz undergo an interaural phase shift of less than 180° when traveling from one ear to the other, and the *interaural phase difference* can uniquely encode the source direction. At higher frequencies, sounds have wavelengths smaller than the head diameter yielding interaural phase differences larger than 180° and thus ambiguous ITDs. Hence, ITDs in stationary high-frequency tones do not provide unique information about the sound's lateral direction. However, in the case of amplitude-modulated and multi-tone sounds, the timing of the envelopes may also be informative even at higher frequencies, yielding envelope ITDs as a useful acoustic feature (Henning 1974).

Yet, broadband ITDs do not provide much information about the sound's spatial position other than its lateral direction. Figure 4b shows iso-ITD contours derived from an exemplary HRTF set. The contours approximate sagittal planes, demonstrating that ITDs are not able to encode the sound's direction on a sagittal plane including the lack of discrimination between front and back. This finding is not new: "*The possibility of distinguishing a voice in front from a voice behind would thus appear to depend on the compound character of the sound in the way that it is not easy to understand, and for which the second ear would be of no advantage*" (Strutt alias Lord Rayleigh 1876). Nowadays, the spatial ambiguity based on the ITD is called the *cone of confusion*, or *torus of confusion* if also distance is involved (Shinn-Cunningham et al. 2000).

3.2.3 Interaural Level Differences

ILDs arise because of two effects. First, the head is an obstacle, creating a shadow for the contralateral ear, and thus ILDs. Depending on the relation between the sound's wavelength and the listener's head size, ILDs increase with both frequency and lateral angle, as shown in Fig. 5a. While low-frequency ILDs span a range of ± 10 dB and increase smoothly for more lateral sounds, high-frequency ILDs exhibit a span of ± 20 dB, with a more complex relation to the lateral angle.

Second, the sound intensity decreases with the distance to the source, creating for near-field sounds, an ILD even at frequencies for which the head is acoustically transparent, as shown in Fig. 5c. Such low-frequency near-field ILDs become significant for distances below 0.5 m (Brungart and Rabinowitz 1999) and can even reach values beyond 20 dB.

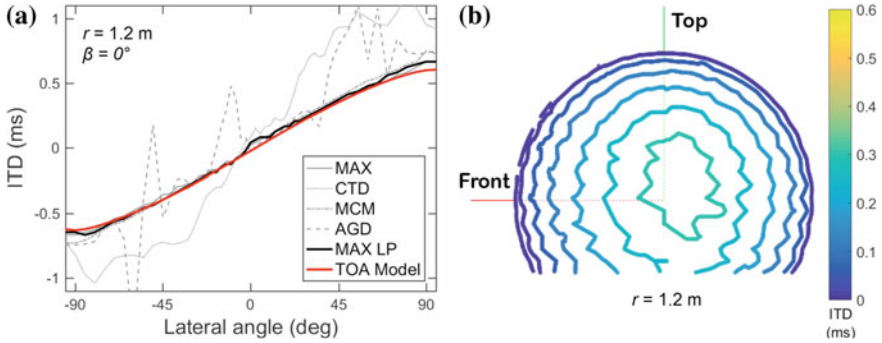


Fig. 4 **a** ITDs for frontal directions at the horizontal plane, estimated by various ITD extraction methods (see text). **b** Iso-ITD contours calculated with the MAX method; view at the left ear along the interaural axis

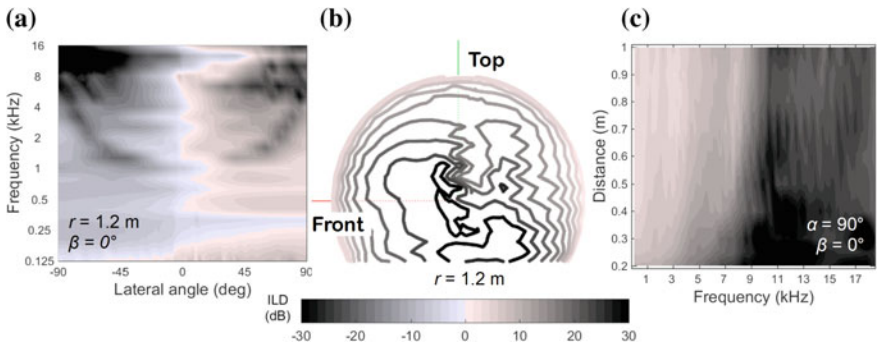


Fig. 5 **a** Frequency-dependent ILDs for sounds along the frontal horizontal half-plane. **b** Broadband iso-ILD contours; view at the left ear along the interaural axis. **c** Frequency-dependent ILDs for the most-left direction in the horizontal plane as a function of distance. Frequency dependence shown by filtering each HRTF by a typical Gammatone filter

Similar to the front-back ambiguity of the ITD, ILDs do not vary consistently with the polar angle, as shown in Fig. 5b. Thus, ILDs do not encode source directions along the sagittal planes well either, further contributing to the cone of confusion based on interaural features.

3.3 Reverberation

So far, only the simplest case of binaural signals was addressed: signals originating from a single sound source in the free field. However, most realistic binaural signals originate from reverberant spaces like rooms. Due to reflections, the binaural signal contains the direct signal overlapped with filtered versions of itself. The filtering

consists of a broadband delay (because of the longer propagation path) and spectral changes (because of the frequency-dependent absorption of the reflecting surfaces and air propagation). Each reflection yields a comb filter in the frequency domain with the spectral density of ripples depending on the delay between the direct and reflected sound.

In realistic situations, often thousands of reflections are created by just a single source. Beginning with the clearly distinguishable early reflections, their temporal density increases such that after some time they become a diffuse field, namely, a sound field with a statistically constant directional distribution. In addition to specular reflections, diffraction and diffusion contribute to the complexity of the reverberant sound field. Acoustically, the effect of reverberation can be described by the *binaural room impulse response* (BRIR), which is basically the binaural pair of HRIRs measured in a room of interest. While a binaural pair of HRIRs are given for the relative relation between the source and listener's position, BRIRs further depend on the absolute positions of the source and listener in the room. As a consequence, in BRIRs, the source-position change is not equivalent to the orientation change of the listener. Hence, a BRIR is a function of at least the source position, listener position, listener orientation, room acoustics, and maybe even source orientation.

The position-dependent effect of reverberation produces binaural signals that are more different than those in free field. Accordingly, the instantaneous interaural similarity in the binaural signal and thus the interaural coherence decreases (Hartmann et al. 2005). For frequencies above 500 Hz, the IACC can even approach zero, depending on the position of the source and listener in a room (Hartmann et al. 2005). The high-frequency envelopes seem to be less susceptible to the reduction of the interaural coherence caused by reverberation as compared to the low-frequency phase differences (Ruggles et al. 2012). Moreover, sound reflections that temporally overlap with the direct sound create instantaneous interaural fluctuations over the time course of the BRIRs, mostly manifesting as a time-dependent IACC. Hence, IACCs calculated over various ranges of time (and frequencies) are widely used in room acoustics (Mason et al. 2005).

3.4 *Dynamic Acoustic Situations*

Even though a sound is an ongoing temporal fluctuation of pressure, the spatial properties of its source do not change unless the spatial configuration between the listener, source, and their environment changes. In spatial hearing, this situation is considered as a *static* one. A widely investigated case is listening to a static sound source without any head movement. In this situation, the HRTFs do not change over time. When the source or listener changes the spatial position and/or orientation, the listening situation becomes *dynamic*: the HRTFs change, creating a systematic temporal change in spatial cues.

In order to describe spatial changes in dynamic listening situations, six degrees of freedom need to be considered for each object with a non-omnidirectional directivity:

three for its orientation and three for its translation. For example, the listener's head orientation can be rotated along the horizontal plane (i.e., *head yaw*), along the median plane (i.e., *head pitch*), and it can be tilted along the frontal plane (i.e., *head roll*). Further, the listener can move along three spatial dimensions (i.e., translation). The same applies to sound sources like musical instruments, talkers, or loudspeakers.

Generally, head pitch changes the orientation of the pinnae relative to the source, causing a change in monaural cues but not necessarily in binaural cues. In contrast, head yaw changes the ears' position in a diametrically opposed fashion, affecting all interaural cues. Hence, horizontal rotations based on head yaw provide dynamic acoustic cues allowing the listener to acoustically resolve the cone of confusion (Perrett and Noble 1995).

The reasons for moving the head due to a sound are manifold. For example, the reflexive orienting response, namely, gaze shift combined with head movements, allows the listener to orient to the source for further inspection (Sokolov 2001). While this reflexive mechanism has been investigated often in the past, not much is known about intentional listeners' head movements in acoustic environments (e.g., Leung et al. 2016). Nevertheless, head movements help in localizing (McAnally and Martin 2014) and externalizing (Brimijoin et al. 2013) sounds as well as tracking auditory targets (Leung et al. 2016)—tasks involved in the formation of 3D auditory space.

4 Neurophysiology: Coding Auditory Space

Auditory spatial perception is formed via neural activities ascending from the auditory nerve to the cortex and is modulated by descending projections, which are not discussed here for simplicity up to the level of the cortex. The ascending auditory pathway can be anatomically separated into the primary pathway and the non-primary pathways (e.g., Straka et al. 2014). Neurons within the primary pathway process mainly auditory information and project in a tonotopic organization to the auditory cortex. The non-primary pathways link a wide constellation of the midbrain, cortical, and limbic-related sites, integrating different types of sensory information and providing information about environmental changes even during sleep. They are not fully understood yet (e.g., Lee 2015). Following the ascending neural organization of the brain, this section describes the contribution of both neural pathways to the formation of the auditory space.

4.1 Subcortical Pathways: Reflexive Map of Auditory Space

Auditory processing begins in the cochlea where inner hair cells produce neural activity, as shown in Fig. 6. The cochlea is not a simple passive sensor; it is an active sensor whose properties are actively modified by the outer hair cells innervated by efferent connections. The inner hair cells transmit the neural information to spiral ganglion cells whose axons form the *auditory nerve* (AN, or cochlear nerve). Each

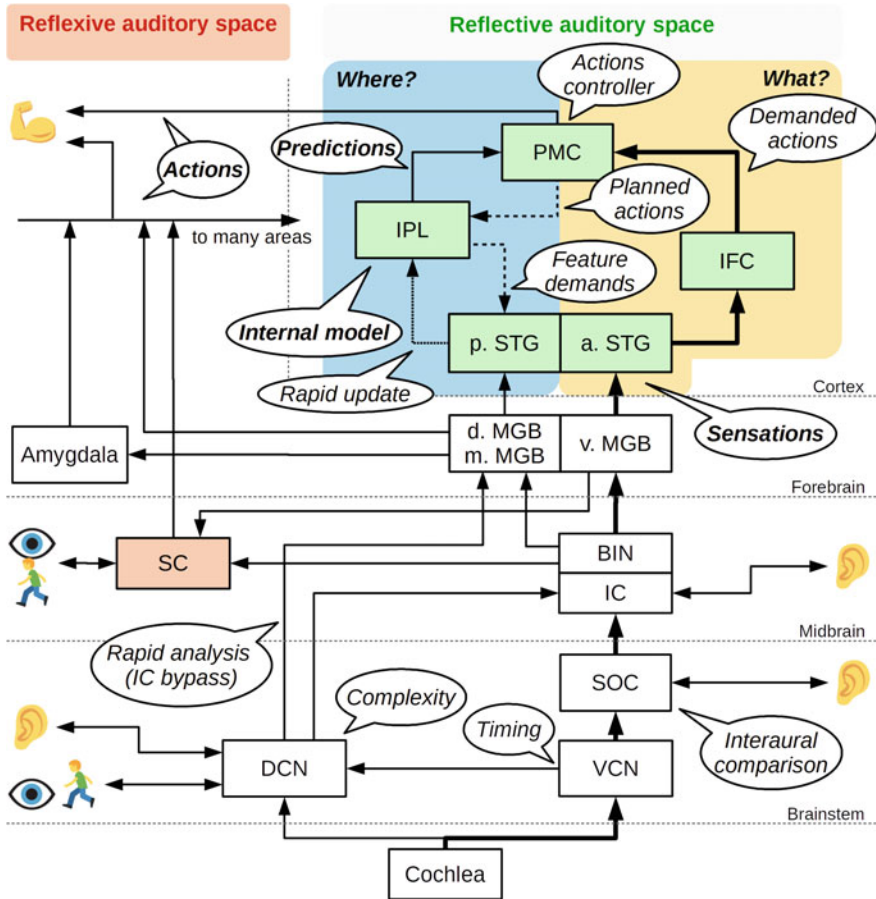






Fig. 6 Ascending neural pathways of human auditory processing (see text for more details). For simplicity, only the unilateral pathway relevant in the formation of spatial representation of the environment, without hemispheric crossings is shown. Bold lines: the primary auditory pathway. Dashed and dotted lines in the “Where?” section: efferent and afferent copies, respectively. Bold text in balloons: correspondence to the cognitive model, as shown in Fig. 1.

 : link to the contralateral pathway (for more details on the binaural processing, see Pecka et al. 2020, this volume);

 : link to the visual system;

 : link to the somatosensory system;

 : link to systems driving actions

of the bilateral ANs projects to the *dorsal cochlear nucleus* (DCN) and the *ventral cochlear nucleus* (VCN) in the medulla of the brainstem. Both nuclei are specialized in decoding certain features of the signal—see Fig. 1 of Dehmel et al. (2008) for more details on this topic.

VCN networks receive input from AN fibers tuned to similar (mostly low) frequencies and comprise bushy and octopus cells, which provide the highest temporal precision of any neuron in the brain. The VCN cells project to the contralateral (via the *trapezoid body*, TB) and ipsilateral *superior olivary complexes* (SOCs). Given its superior timing properties, the VCN provides solid a basis for ITD decoding.

The DCN is organized differently as compared to the VCN. Its cells receive inputs not only from the high-frequency AN fibers but also from the VCN and various efferent types of sensory circuits (somatosensory, reticular, vestibular; Muniak and Ryugo 2014). They are similar to those found in the cerebellum, providing evidence for their capability of complex information processing. Indeed, they form tonotopically organized networks that perform nonlinear spectral analysis and separate spectral features according to their expectancy (Singla et al. 2017). This analysis helps in forming acoustic cues required for sound-source localization in sagittal planes (May 2000). Further, it may play a role in attenuating body-generated sounds such as vocalization or respiration and thus may improve the signal-to-noise ratio of the received signal (Shore 2005).

Interestingly, DCN cells also receive (mostly inhibitory) inputs from the contralateral CN, indicating integration of binaural information already at this very early level of neural processing. The projections of the DCN cells are manifold. Many of them go directly to the *inferior colliculus* (IC), conveying auditory features including the information required for sound localization based on monaural spectral cues. Other projections are more spread and include a direct path to the thalamus, bypassing the IC, allowing the cortex to prepare for rapid analysis (Anderson et al. 2006), and a path to the nucleus reticularis pontis caudalis, critical for triggering the acoustic startle reflex (Meloni and Davis 1998).

In the SOC, information is mostly processed in three primary nuclei: the *medial superior olive* (MSO), the *lateral superior olive* (LSO), and the medial and lateral nuclei of the TB. The MSO and LSO have a tonotopic organization with bilateral inputs from the VCNs. While the MSO decodes the interaural phase differences (IPDs), which represent the ITDs, the LSO is mostly associated with the decoding of the ILDs (see Pecka et al. 2020, this volume). Combined with the DCN, basic auditory spatial features like the ITD, ILD, and monaural spectral features are partially decoded at this level of the neural pathway, forming spatial cues available for processing in further neural structures relevant to auditory space.

In the midbrain, the IC is an obligatory relay for most of the ascending auditory information, and it combines information from other modalities (Gruters and Groh 2012). Acoustic features are prepared for the formation of auditory objects happening at the next synaptic level. The IC is divided into at least three parts. The central nucleus of the IC is exclusively for auditory tasks. It is organized in sheets of isofrequency laminae, each of which receives inputs from multiple different nuclei of the brainstem that permit the decoding of parallel attributes like amplitude and frequency modulation. Its binaural interactions appear to be quite complex and nonlinear. Nevertheless, there is strong evidence for the processing of spatial information in all three dimensions. Single neurons have been found showing ITD sensitivity similar to that found in psychophysical experiments. These neurons enable the decoding of the lat-

eral angle of a sound using a single neural path without the need for a population code (Skottun et al. 2001). Single neurons have also been found to demonstrate better temporal coding of reverberant stimuli than that of anechoic stimuli (Slama and Delgutte 2015). This form of de-reverberation of the signals may not only improve robustness in sound recognition but may also be used to estimate the distance of a sound source by enabling comparisons between direct and reverberant sound energy. Finally, neurons sensitive to simple analogs of HRTF-like spectral shapes have been found (Davis et al. 2003), which show the presence of cells sensitive to the spectral cues encoding the polar angle of the sound source. The pericentral and external parts of the IC receive ascending inputs from the central IC as well as descending inputs from the somatosensory system, auditory cortex, and other higher brain regions. Many projections are bilateral and thus create numerous feedback loops enabling the integration of the processed auditory information with that arriving from other sensory systems.

From the IC, the information is transmitted via the *brachium of the inferior colliculus* (BIN) in parallel to the thalamus and *superior colliculus* (SC). The BIN seems to be involved in processing the spatial auditory information where some neurons seem to prefer the natural alignment of the interaural and spectral spatial cues provided by a realistic sound of an acoustic source (Slee and Young 2014).

The SC integrates information from multiple modalities, in particular, auditory spatial with visual and somatosensory information. In the SC, maps of the visual space, body surface, and auditory space arise from spatially ordered projections from the retina, skin, and acoustic features. Interestingly, ITDs do not seem to contribute much to that spatial map. Instead, spatial tuning seems to mostly rely on spectral ILDs (Slee and Young 2013). Already at this early neural level, the cooperation of neurons combining multimodal afferent and efferent projections decreases reaction time, increases stimulus detectability, and enhances perceptual reliability when information from two modalities is required to accomplish behavioral tasks (Kayser and Logothetis 2007). The SC projects to many motor-related parts of the brain and its organization can be seen as a dynamic map of motor error (Middlebrooks 2009, pp. 745) with receptive fields reflecting the deviation between the angle of gaze (including head and eye position) and the target defined by the sensory visual, auditory, and somatic inputs. As a whole, the SC plays a critical role in the ability to direct behaviors toward specific objects by orienting the head and eyes towards something seen and heard (Klier 2003).

Interestingly, on the one hand, these neural and behavioral abilities seem to remain active even in the absence of the cortex (Woods 1964). On the other hand, dark-reared animals showed a lack of the SC auditory map while still being able to perform auditory-based behavioral tasks (Blatt et al. 1998). Although, the SC demonstrates the formation of a neural topographic map of the auditory space, the SC is most likely not responsible for the creation of the conscious auditory space. Instead, it is a quick and reflexive way to react to the environment in the form of an orienting response (Peck 1996). From this observation, one might conclude that the pathway from SOC via IC to SC forms a reflexive map of the auditory space (see also Yao et al. 2015), in parallel to the primary ascending path from the IC to the thalamus and cortex.

4.2 *Thalamus and Cortex: Cognition of Auditory Space*

The cortex plays an important role in the formation of auditory space because it groups and segregates spatial and non-spatial features of the sensory information into streams referring to auditory objects with specific properties—see Micheyl et al. (2007) and van der Heijden et al. (2019) for a review. Before reaching the cortex, the *thalamus* acts as a hub, relaying information between different subcortical areas and the cerebral cortex. Also, the thalamus has been shown to provide a critical integration of other modalities and the preparation of motor responses. Within the thalamus, the auditory information is processed by the *medial geniculate body* (MGB) of the ventral thalamus. The MGB is further split into dorsal, medial, and ventral sections.

The *dorsal* MGB is organized non-tonotopically. Its auditory responses can be influenced strongly by non-auditory inputs, and its neurons project mainly to the belt of the auditory cortex (Bartlett 2013). The *medial* MGB receives various inputs from the BIN and projects broadly across many tonotopic, non-tonotopic, multimodal, and limbic cortical areas, terminating in the cortex and amygdala—see Lee (2015) for a review on this topic. In addition, a direct pathway from the DCN to the medial MGB bypasses the IC (Anderson et al. 2006). This connection shows lower latencies than those via the IC, which are advantageous for priming the auditory cortex to be prepared for rapid analysis and for recruiting the amygdala for rapid emotional responses such as fear. This rapid emotional analysis potentially triggers the startle reflex and auditory looming bias (Bach et al. 2008). It is not clear yet how the dorsal and medial parts of the MGB contribute to the process of forming auditory objects.

The *ventral* MGB has been investigated more thoroughly. It processes mostly auditory information, it is driven by projections from the central IC, and it forms a tonotopically organized relay of binaural (most) and contralateral-only (little) information (Lee and Sherman 2010). Its outputs mostly project to the primary auditory cortex and the rostral auditory area where the acoustic features are further processed to build auditory objects.

When finally reaching the cortex, the auditory information is spread among various regions. The most affected regions are the *superior temporal gyrus* (STG, including the *auditory cortex*, AC), *inferior frontal cortex* (IFC), *inferior parietal lobe* (IPL), and *premotor cortex* (PMC), as shown in Fig. 6. There is strong evidence for the existence of two largely segregated processing paths, forming the dual-pathway model of auditory cortical processing, each of them subserving the two main functions of hearing: “what”, the identification of auditory objects (e.g., recognition of talkers); and “where”, the processing of motion and spatial properties of objects (Rauschecker and Tian 2000).

The “what” pathway follows the anterolateral route of the AC, which includes the *primary auditory cortex* (A1), a rostral area, the lateral and medial belt (including the caudomedial area), and the parabelt. Moving along the ascending route within the A1, a transition from the representation of acoustic features (e.g., response to pure tones) via perceptual features (e.g., pitch and timbre) to category representation (e.g., auditory objects) happens. Beginning in the A1, pure-tone sensitive neurons

receive inputs from the ventral MBG and are mostly tonotopically organized. A topographic representation of the auditory space, like that found in the SC, is mostly missing. In contrast, responses of some single A1 neurons result in a 360°-like panoramic representation of space (Middlebrooks 2009). The “what” pathway ascends further to the IFC via other processing areas within the anterolateral STG. These and similar connections enable further processing of non-spatial properties of auditory objects. From the IFC, the auditory information is transformed into articulatory or motor representations in the PMC.

In contrast, the “where” pathway is highly involved in processing the *spatial* information retrieved from the auditory stream—see van der Heijden et al. (2019) for a review on this topic. While its exact role is still being debated, its processing involves the separation, location, trajectory, and temporal context of a sound (Ahveninen et al. 2014). The active regions of the “where” pathway are quite diverse; they seem to follow a posterodorsal non-primary route, with projections from medial and dorsal MGB via the posterior STG (including the planum temporale, i.e., the superior surface of the STG) to the IPL and dorsal and ventral areas in the PMC (Rauschecker and Tian 2000). The spatial information, including ITD and binaural coherence, can be found encoded by a population code in various areas of the posterior STG (Miller and Recanzone 2009). The spatial information is further processed by the IPL (Arnott et al. 2004) where spatial features are integrated with information from other sensory modalities and further projected to the PMC.

Hence, the PMC is activated by the “what” pathway via the IFC, as shown by the bold lines in Fig. 6, and it is also modulated by the “where” pathway via the IPL, as shown by the blueish section of Fig. 6. This modulation corresponds to a feedforward system consisting of an internal predictive model of the environment (located in the PMC) updated by the multimodal sensory information (ascending from the IPL). This allows the motor system to quickly adapt to the new sensory situation.

The PMC’s ability to react to predictions based on sensory information is further underlined by findings showing that the PMC is not only involved during acoustic stimulation but also during musical imagery (Leaver et al. 2009). To this end, the PMC assembles motor patterns for the potential production of sound sequences. Efferent feedback from the PMC about planned motor actions (efferent copy; dashed line from the PMC in Fig. 6), together with the fast and temporally precise afferent projections from the posterior STG (afferent copy; dotted line to the ILP in Fig. 6) allow the IPL to compare the spatial auditory information with the predicted motor states, to decide about the required adjustments of the internal model, and to minimize the surprise—compare Sec. 2.2. Further efferent projections to the STG (dashed line in Fig. 6) modulate the process of feature extraction in the STG according to the changing feature demands—see Rauschecker (2011) for more details.

These considerations show that many cortical regions are involved in processing features of the auditory space and there is no clear evidence for a single region representing the auditory space in the cortex per se. The spatial information is represented via the firing rate of the neural population. This process is further modulated by vision (e.g., Mendonça 2020, this volume), proprioception (Genzel et al. 2016), and attention (e.g., Deng et al. 2019, and Fels et al. 2020, this volume), indicating

its reflective nature in the formation of the auditory space. This is in contrast to the well-localized but reflexive map of the auditory space found at the level of the SC.

The summary presented in this section is just a simplification of all the reflexive and reflective processes involved. The human brain is an analog, high-dimensional, recurrent, nonlinear, stochastic, and dynamic system (Dotov 2014). Together, these processes form the perception that allows humans to complete various spatial tasks. The following section describes psychophysical *spatial* tasks, all of them demonstrating the ability to utilize the understanding of the 3D auditory space.

5 Psychophysics: Listener's Abilities Given the Perceived Auditory Space

Spatial auditory cues facilitate both reflexive and reflective behavior. Human psychoacoustic studies usually imply reflective behavioral tasks, while reflexive behavior, if not targeted explicitly, is more commonly tested in populations with limited cognitive abilities such as infants. This section reviews reflective spatial abilities, ordered by their increasing cognitive complexity.

5.1 Sound Localization

Sound localization describes the (reflective) ability to estimate the spatial position of the sound source (Middlebrooks 2015). Sound localization experiments are often conducted in anechoic environments or in a virtual space simulating a free field. Target sounds are typically presented via loudspeakers or headphones. For the latter, the auditory space is often simulated by filtering the sounds with HRTFs. The cues contributing to sound localization are conceptually different for each dimension of the interaural-polar coordinate system.

Interaural cues facilitate sound localization within the lateral dimension. The duplex theory describes ITD cues being most useful for low-frequency sounds and ILD cues for high-frequency sounds, however, localization of broadband sounds is dominated by low-frequency ITDs (Macpherson and Middlebrooks 2002). Sounds can be accurately localized depending on their lateral angle. Horizontal *minimum audible angles* (MAAs) between two successively presented sounds can be as small as 1° for frontal sources but increase with lateral angles to approximately 10° (Perrott and Saberi 1990). These thresholds were obtained in free-field experiments using natural combinations of ITD and ILD cues. Tested in isolation, discrimination thresholds of IPDs (linked to ITDs) also increase by an order of magnitude with increasing lateral reference angles (e.g., 2° at 0° reference increasing to 20° at 90° reference for a 900 Hz tone, Yost 1974), closely resembling the observed MAAs. ITD cues result from the temporal fine structure at frequencies below approximately 1.4 kHz

and from the temporal envelopes at higher frequencies. Consistent with their small perceptual weight in the duplex theory, *just-noticeable differences* (JNDs) for such high-frequency envelope ITDs are at least twice as large as the JNDs at low frequencies (Bernstein and Trahiotis 2002). ILD JNDs are in the range of 1 dB for high-frequency sounds and depend on amplitude modulation rates. Effects of temporal modulations on ILD JNDs can be explained based on the interaural difference in neural discharge rates with no need for a particular binaural adaptation mechanism (Laback et al. 2017). Computational models of lateral-angle localization have a long history and diversity. Recently, a large-scale attempt to systematically investigate these approaches has been initiated (Dietz et al. 2017). The auditory system is at least partly able to adapt to changes in ITD and ILD cues according to visual and audiomotor feedback (e.g., Trapeau and Schönwiesner 2015). Adaptation after-effects were only observed in short-term but not long-term studies (Trapeau and Schönwiesner 2015), indicating that different mechanisms are involved.

Spectral-shape cues are crucial for sound localization within the sagittal dimension. For lateral-angle localization, they are important in monaural hearing only (Macpherson and Middlebrooks 2002). Consequently, in order to achieve high spatial acuity as indicated by vertical MAAs as small as 4° , localization in sagittal planes requires a bandwidth of 16 kHz (Perrott and Saberi 1990) and limitations down to 8 kHz cause marked degradations (Best et al. 2005). Spectral degradations often result in localization responses biased toward the horizon and led to the concept of *elevation gain* in polar-angle localization (Hofman et al. 1998). Spectral-shape cues are arguably processed within monaural pathways and thus are often referred to as “monaural spectral cues” although information from both ears is combined following a spatially systematic binaural weighting scheme (Macpherson and Sabin 2007): The contralateral ear contributes less with increasing lateral eccentricity. Because of monaural processing, localization performance can be affected by frequency modulations in the stimulus spectrum. Template-based models can explain these interactions and show how monaural spectral cues, when extracted based on tonotopic gradients, are rather independent of naturally appearing low-frequency spectral modulations in the source spectrum (Baumgartner et al. 2014). Localization performance along sagittal planes is particularly listener-specific, but this variation is hardly explained by only considering the acoustic factor of listener-specific HRTFs, suggesting large inter-individual differences in how efficient the auditory system can utilize the acoustic information (Majdak et al. 2014; Baumgartner et al. 2016). Further, listeners are able to learn new spectral-shape cues (Hofman et al. 1998; Majdak et al. 2013b) and even use them simultaneously with previously acquired cues (Trapeau et al. 2016).

In order to estimate the distance of a source, listeners have access to a broad variety of acoustic cues like sound intensity, reverberation characteristics (often quantified by the direct-to-reverberant energy ratio), near-field ILDs, the shape of the stimulus spectrum, and others—see Kolarik et al. (2016) for a review. The relative perceptual relevance of these cues and their underlying neural codes are the subject of debate and are most probably dependent on the context—see HlÁdek et al. (2017) for a review on this topic. Recent studies suggest that the amount of temporal ILD fluctuations and amplitude modulation depth likely represent reverberation-related cues (Catic et al.

2015). Moreover, modifications of spectral-shape cues can affect distance perception (Baumgartner et al. 2017) and familiarization to non-individualized spectral-shape cues can improve distance perception (Mendonça et al. 2013).

A special case of distance perception concerns distances closer than physically plausible, namely, inside the listener's head. Sounds are naturally perceived outside the head (externalized) whereas sounds reproduced with headphones or hearing-assistive devices are often perceived internally and appear to originate from inside the head (Noble and Gatehouse 2006). *Sound externalization* is not directly related to the playback device, as free-field signals can be internalized as well (Brimijoin et al. 2013). Sound externalization has been investigated by means of discrimination tasks between real and virtual sources and/or distance ratings (Hartmann and Wittenberg 1996). Although perceived distance may be only one of many cues to discriminate between virtual and real sources, the similarity of findings for both paradigms suggests that distance perception is somehow linked with sound externalization. One could think that the only reason for using the term “sound externalization” instead of “distance” is that the percept of sound internalization is an available option. At first glance, however, there seem to be some contradictions between studies focusing on sound externalization and distance perception. For example, decreasing low-frequency ILDs were associated with increasing distance (Brungart et al. 1999), whereas ILDs gradually removed from the low-frequency partials of a harmonic complex (Hartmann and Wittenberg 1996) or decreased from broadband speech (Brimijoin et al. 2013) were associated with reduced sound externalization. One might conclude that sound internalization is a default state for the case of missing (or implausible) auditory cues, for which no plausible internal model of the environment can be established.

In addition to the static sound localization, self motion introduces dynamic cues to the binaural signal. This interaction is successfully compensated by mechanisms responsible for building an allocentric frame of reference (Yost et al. 2015). Listeners are able to create such allocentric spatial representations even without visual information (Viaud-Delmon and Warusfel 2014) and such representations help in resolving front-back confusions (McAnally and Martin 2014). Source motion also modifies the binaural signal, but, in contrary to self-motion, source motion does not require active actions of the listener. Listeners are able to detect source-movement angles as small as 2° , depending on source velocity, stimulus duration, and bandwidth—see Carlile and Leung (2016) for more details on this topic.

Finally, the contribution of the cues to the process of sound localization is not static. Alterations of the acoustic environment may change the informative character of a spatial cue and consequently affect its contribution to the process of localization (Keating et al. 2015). Neural plasticity not only enables such context-dependent re-weighting of the cues, but also enables adaptation to a new set of spatial cues—see Mendonça (2014) for a detailed review.

5.2 From Spatial Impression to Presence

In reverberant spaces like rooms or concert halls, the auditory perception becomes multidimensional (Cerdá et al. 2009). Listeners are still able to localize the direct sound even in the presence of early reflections, although such reflections technically might represent separate auditory objects. The ability to suppress early reflections and actually to not perceive them as echoes is referred to as the precedence effect—see Clapp and Seeber (2020), this volume, for review. The delay of the reflections relevant for the precedence effect depends on the stimulus; it is around 5 ms for short clicks and can be up to 30 ms for complex stimuli like speech. The presence of reverberation, however, introduces other spatial effects, which have been summarized as *spatial impression* or *spaciousness* (Kuhl 1978). They consist of two main components: *apparent source width* (ASW) and *listener envelopment* (LEV, Bradley and Soulodre 1995).

The ASW describes the spatial width of a sound event perceived by the listener. In headphone experiments, the primary cue for the compactness of the perceived sound is the IACC. If it decreases, the sound is perceived as a wider image (Blauert and Lindemann 1986). Interestingly, for narrow-band sounds, listeners are extremely sensitive to the deviation of a perfectly coherent signal; they can easily discriminate between signals with an IACC of 1 and 0.99 (Gabriel and Colburn 1981). The current explanation for such high sensitivity is that even in a slightly incoherent signal, large instantaneous ITD and ILD fluctuations occur, which can easily be detected by the auditory system. The perceptual consequence of the fluctuations depends on their duration, bandwidth, center frequency, and stimulus sound level (Goupell and Hartmann 2006). The ASW gradually declines with increasing IACC, but it is hardly affected between IACCs of 1 and 0.99 (Whitmer et al. 2013). For the extreme case of interaural decorrelation (IACC of zero), the perceived auditory image splits into two objects appearing in the left and the right ears, respectively.

In reverberant environments where multiple reflections overlap the direct sound, the ASW is determined by the mid-frequency lateral energy fraction and IACC of the early arriving sound field, that is, within the first 80 ms of the BRIR (Okano et al. 1998). Deviations of that IACC from 1 contributes to the perceived quality of concert halls and has been termed the *binaural quality index* (BQI) of room acoustics. Interestingly, as the BQI increases, the low-frequency ITDs are more likely to be disturbed. As a consequence, ITD-based spatial hearing in reverberant situations seems to rely more on high-frequency ITD cues (transmitted in the signal envelope) than on low-frequency ITD cues (Ruggles et al. 2012).

LEV describes how much the listener feels to be immersed in the sound field. LEV depends on the level, direction of arrival, and temporal properties of later (after 80 ms) arriving reflections (Bradley and Soulodre 1995). Late sound arriving from the side, overhead, and behind the listener correlates strongly with LEV. LEV can be distinguished from the late sound having non-lateral components (Furuya et al. 2001) and the late sound arriving from behind and above the listener seems to be important as well (Morimoto et al. 2001), together suggesting the perception of rooms relies on an accurate formation of 3D auditory space.

The ASW and LEV can be predicted with a model based on the BRIRs and JNDs of the IACC (Klockgether and van de Par 2014). In order to quantitatively explain the overall perceived quality of concert halls, additional consideration of the reverberation time is required (Cerdá et al. 2015).

While the ASW and LEV seem to be the major parameters describing the spatial impression of a room, they can both be seen as parts of a broader concept widely used in the context of *virtual environments*. Here, immersion, as a measure of the psychological sensation of being surrounded (Begault et al. 1998), integrates the objectively derived LEV and ASW, and further extends to “*a psychological state characterized by perceiving oneself to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli and experiences*” (Witmer and Singer 1998). In that context, responses to a given level of immersion have been defined as *presence*, a measure of the psychological sensation of being elsewhere (e.g., Slater 2003). Both immersion and presence are essential for the quality of experience in virtual environments (Möller and Raake 2014). For example, in headphone-based virtual environments, immersion can be enhanced with the use of listener-specific HRTFs (Wenzel et al. 1990; Blauert et al. 2000; Djelani et al. 2000; Vorländer and Shinn-Cunningham 2014), being in line with neurocognitive and neurophysiological studies showing that the auditory system prefers the natural combination of spatial cues (Deng et al. 2019; Salminen et al. 2015; Slee and Young 2014). Interestingly, immersion seems to be more easily conveyed via audio than vision because audio operates all around the listener even outside the listener’s field of view and without exploratory head movements. Immersion and presence seem to be highly related attributes, and the underlying mechanisms are not fully understood yet—see Gaggioli et al. (2003) for a review.

5.3 Other Spatially-Related Tasks

Spatial hearing also improves tasks not directly related to the formation of the 3D space. A famous example is the *cocktail-party effect* that describes the ability to focus on and thus to improve the intelligibility of a particular talker in a multi-talker environment (Cherry 1953; Bronkhorst 2015). In such a task, the formation of the spatial world is not required per se, however, the benefit of spatial separation of maskers from the target, also called *spatial unmasking* or *spatial release from masking*, is clear and has been considered in models predicting speech intelligibility from binaural signals in many situations (e.g., Lavandier et al. 2012). Spatial unmasking can further reduce cognitive load in conditions providing similar speech intelligibility (e.g., Andéol et al. 2017). Spatial attention, that is, knowing “where” to focus, further modulates the effect of spatial unmasking on a very listener-specific basis (Oberfeld and Klöckner-Nowotny 2016).

Note that spatial unmasking is not only limited to spatially separated targets and/or noise. Improved speech intelligibility has also been shown in listeners once they have adapted to the acoustics of the listening room (Brandewie and Zahorik 2010) indicating that while the auditory system can adapt to reverberant spaces, the

“masker” in spatial unmasking can be both an additional sound source and acoustic reflections of the same source.

Spatial hearing contributes to other, less-known non-spatial tasks. For example, spatial impression can increase the emotional impact of orchestra music by enhancing musical dynamics—see Pätynen and Lokki (2016) and Lokki and Pätynen (2020). Looming bias, namely, the phenomenon that approaching sounds are more salient than receding sounds, can be mediated significantly by sound externalization created by the acoustic spatial pinna features alone (Baumgartner et al. 2017). These and similar findings underline the relevance of the formation of 3D auditory space in people’s everyday life.

6 Conclusions

The formation of the auditory space is one of the cognitive processes required to understand and interact with the environment. In that process, the auditory system has to cope with ephemeral acoustic information created by objects surrounding the listener. Their spatial information, conveyed by the binaural signals, is encoded by interaural and monaural features along various temporal ranges. The neural auditory system then creates two representations of the auditory space: a topographically structured neural network in the superior colliculus, capable of triggering quick reflexive reactions; and a reflective cortical representation, encoded by neural populations capable of modulating other cognitive processes through attention. The reflective representation allows humans to perceive the auditory space and consciously perform spatial tasks.

Many concepts have been proposed for cognitive processes involved in the formation of the auditory space. The listener’s interaction with the environment can be seen as a feedforward system with an internal model predicting the external (or distal) state of affairs. Feedback coming from the auditory and other senses allows the listener to compensate for any deviations to the predictions. Given the ambiguity in the estimation of the external state of affairs from the limited binaural information, the free-energy principle and the active inference seem to be promising approaches to explain how cognition restricts itself to a limited number of plausible states. Further progress in the development of mathematical methods for solving ill-posed problems and of experimental methods combining psychophysics with neurophysiology will help to improve the understanding of the formation of the auditory space in the future. This is a prerequisite for advances in many technical applications like hearing aids driven by spatial attention, listener-specific virtual acoustics, and dynamic sound reproduction systems.

Acknowledgements We thank S. Clapp and B. Seeber as well as two anonymous reviewers for their valuable comments and suggestions. The work presented in this chapter was supported by the Austrian Science Fund (FWF, J 3803-N30) and the European Commission (Project ALT, Grant 691229).

References

- AES69-2015. 2015. AES standard for file exchange—Spatial acoustic data file format Standard .
- Ahveninen, J., N. Kopčo, and I.P. Jääskeläinen. 2014. Psychophysics and neuronal bases of sound localization in humans. *Hearing Research* 307: 86–97. <https://doi.org/10.1016/j.heares.2013.07.008>.
- Aitchison, L., and M. Lengyel. 2017. With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology* 46: 219–227. <https://doi.org/10.1016/j.comb.2017.08.010>.
- Algazi, V.R., C. Avendano, and R.O. Duda. 2001a. Elevation localization and head-related transfer function analysis at low frequencies. *Journal of the Acoustical Society of America* 109 (3): 1110–1122. <http://view.ncbi.nlm.nih.gov/pubmed/11303925>.
- Algazi, V.R., C. Avendano, and R.O. Duda. 2001b. Estimation of a spherical-head model from anthropometry. *Journal of the Audio Engineering Society* 49 (6): 472–479. <http://www.aes.org/e-lib/browse.cfm?elib=10188>.
- Andéol, G., C. Suied, S. Scannella, and F. Dehais. 2017. The spatial release of cognitive load in cocktail party is determined by the relative levels of the talkers. *Journal of the Association for Research in Otolaryngology* 1–8: <https://doi.org/10.1007/s10162-016-0611-7>.
- Anderson, L.A., M.S. Malmierca, M.N. Wallace, and A.R. Palmer. 2006. Evidence for a direct, short latency projection from the dorsal cochlear nucleus to the auditory thalamus in the guinea pig. *European Journal of Neuroscience* 24 (2): 491–498. <https://doi.org/10.1111/j.1460-9568.2006.04930.x>.
- Andreopoulou, A., and B.F.G. Katz. 2017. Identification of perceptually relevant methods of interaural time difference estimation. *Journal of the Acoustical Society of America* 142 (2): 588–598. <https://doi.org/10.1121/1.4996457>.
- Arnal, L.H., A. Flinker, A. Kleinschmidt, A.-L. Giraud, and D. Poeppel. 2015. Human screams occupy a privileged niche in the communication soundscape. *Current Biology* 25 (15): 2051–2056. <https://doi.org/10.1016/j.cub.2015.06.043>.
- Arnott, S.R., M.A. Binns, C.L. Grady, and C. Alain. 2004. Assessing the auditory dual-pathway model in humans. *NeuroImage* 22 (1): 401–408. <https://doi.org/10.1016/j.neuroimage.2004.01.014>.
- Awh, E., A.V. Belopolsky, and J. Theeuwes. 2012. Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences* 16 (8): 437–443. <https://doi.org/10.1016/j.tics.2012.06.010>.
- Bach, D.R., H. Schächinger, J.G. Neuhoff, F. Esposito, F. Di Salle, C. Lehmann, M. Herdener, K. Scheffler, and E. Seifritz. 2008. Rising sound intensity: An intrinsic warning cue activating the amygdala. *Cerebral Cortex* 18 (1): 145–150. <https://doi.org/10.1093/cercor/bhm040>.
- Bartlett, E.L. 2013. The organization and physiology of the auditory thalamus and its role in processing acoustic features important for speech perception. *Brain and Language* 126 (1): 29–48. <https://doi.org/10.1016/j.bandl.2013.03.003>.
- Baumgartner, R., Majdak, P., and Laback, B. 2016. Modeling the effects of sensorineural hearing loss on sound localization in the median plane. *Trends HearTrends in Hearing* 20: 2331216516662003. <https://doi.org/10.1177/2331216516662003>.
- Baumgartner, R., P. Majdak, and B. Laback. 2014. Modeling sound-source localization in sagittal planes for human listeners. *Journal of the Acoustical Society of America* 136 (2): 791–802. <https://doi.org/10.1121/1.4887447>.
- Baumgartner, R., D.K. Reed, B. Tóth, V. Best, P. Majdak, H.S. Colburn, and B. Shinn-Cunningham. 2017. Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias. *Proceedings of the National Academy of Sciences of the United States of America* 114 (36): 9743–9748. <https://doi.org/10.1073/pnas.1703247114>.
- Begault, D.R., Ellis, S.R., and Wenzel, E.M. 1998. Headphone and head-mounted visual displays for virtual environments. *Journal of the Audio Engineering Society*, 49: 904–916.

- Bernstein, L.R., and C. Trahiotis. 2002. Enhancing sensitivity to interaural delays at high frequencies by using transposed stimuli. *Journal of the Acoustical Society of America* 112 (3): 1026–1036.
- Best, V., S. Carlile, C. Jin, and A. van Schaik. 2005. The role of high frequencies in speech localization. *Journal of the Acoustical Society of America* 118 (1): 353–363. <https://doi.org/10.1121/1.1926107>.
- Best, V., F.J. Gallun, S. Carlile, and B.G. Shinn-Cunningham. 2007. Binaural interference and auditory grouping. *Journal of the Acoustical Society of America* 121 (2): 1070–1076.
- Bhatt, R.S., and P.C. Quinn. 2011. How does learning impact development in infancy? The case of perceptual organization. *Infancy* 16 (1): 2–38. <https://doi.org/10.1111/j.1532-7078.2010.00048.x>.
- Bizley, J.K., and Y.E. Cohen. 2013. The what, where and how of auditory-object perception. *Nature Reviews Neuroscience* 14 (10): 693–707. <https://doi.org/10.1038/nrn3565>.
- Blatt, B., E. von Linstow Roloff, D.J. Withington, E.M. Macphail, and G. Riedel. 1998. Analysis of the superior colliculus auditory space map function in guinea pig behavior. *Neuroscience Research Communications* 23 (1): 23–40.
- Blauert, J. 1997. *Spatial Hearing. The Psychophysics of Human Sound Localization*. Cambridge, MA: The MIT Press.
- Blauert, J. 1969. Sound localization in the median plane. *Acustica* 22: 205–213.
- Blauert, J., and G. Brown. 2020. Reflexive and reflective auditory feedback. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 3–32. Cham, Switzerland: Springer and ASA Press.
- Blauert, J., and W. Lindemann. 1986. Spatial mapping of intracranial auditory events for various degrees of interaural coherence. *Journal of the Acoustical Society of America* 79 (3): 806–813. <https://doi.org/10.1121/1.393471>.
- Blauert, J., H. Lehnert, J. Sahrhage, and H. Strauss. 2000. An interactive virtual-environment generator for psychoacoustic research. I: Architecture and implementation. *Acta Acustica United with Acustica* 86 (1): 94–102.
- Bradley, J.S., and G.A. Soulodre. 1995. Objective measures of listener envelopment. *Journal of the Acoustical Society of America* 98 (5): 2590–2597. <https://doi.org/10.1121/1.413225>.
- Brandewie, E., and P. Zahorik. 2010. Prior listening in rooms improves speech intelligibility. *Journal of the Acoustical Society of America* 128 (1): 291–299. <https://doi.org/10.1121/1.3436565>.
- Bregman, A.S. 1990. *Auditory Scene Analysis, 10*. Cambridge, MA: MIT Press.
- Brimijoin, W.O., A.W. Boyd, and M.A. Akeroyd. 2013. The contribution of head movement to the externalization and internalization of sounds. *PLoS One* 8 (12): e83068. <https://doi.org/10.1371/journal.pone.0083068>.
- Bronkhorst, A.W. 2015. The cocktail-party problem revisited: Early processing and selection of multi-talker speech. *Atten Percept Psychophys* 77 (5): 1465–1487. <https://doi.org/10.3758/s13414-015-0882-9>.
- Brungart, D.S., and W.M. Rabinowitz. 1999. Auditory localization of nearby sources. Head-related transfer functions. *Journal of the Acoustical Society of America* 106 (3): 1465–1479.
- Brungart, D.S., N.I. Durlach, and W.M. Rabinowitz. 1999. Auditory localization of nearby sources. II. Localization of a broadband source. *Journal of the Acoustical Society of America* 106 (4): 1956–1968. <https://doi.org/10.1121/1.427943>.
- Carbon, C.-C. 2014. Understanding human perception by human-made illusions. *Frontiers in Human Neuroscience* 8: 566. <https://doi.org/10.3389/fnhum.2014.00566>.
- Carlile, S., and C. Corkhill. 2015. Selective spatial attention modulates bottom-up informational masking of speech. *Scientific Reports* 5 (1): 8662. <https://doi.org/10.1038/srep08662>.
- Carlile, S., and J. Leung. 2016. The perception of auditory motion. *Trends Hearing* 20: 1–19. <https://doi.org/10.1177/2331216516644254>.
- Catic, J., S. Santurette, and T. Dau. 2015. The role of reverberation-related binaural cues in the externalization of speech. *Journal of the Acoustical Society of America* 138 (2): 1154–1167. <https://doi.org/10.1121/1.4928132>.

- Cerdá, S., A. Giménez, J. Romero, R. Cibrián, and J. Miralles. 2009. Room acoustical parameters: A factor analysis approach. *Applied Acoustics* 70 (1): 97–109. <https://doi.org/10.1016/j.apacoust.2008.01.001>.
- Cerdá, S., A. Giménez, R. Cibrián, S. Girón, and T. Zamarreño. 2015. Subjective ranking of concert halls substantiated through orthogonal objective parameters. *Journal of the Acoustical Society of America* 137 (2): 580–584. <https://doi.org/10.1121/1.4906263>.
- Chandrasekaran, B., Koslov, S.R., and Maddox, W.T. 2014. Toward a dual-learning systems model of speech category learning. *Frontiers in Psychology* 5: 825. <https://doi.org/10.3389/fpsyg.2014.00825>.
- Cherry, E.C. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America* 25 (5): 975–979.
- Clapp, S., and U.B. Seeber. 2020. Auditory room learning and adaptation to sound reflections. In *The Technology of Binaural Understanding*, eds. by Blauert, J. and Braasch, J. Springer and ASA Press. This volume.
- Curtis, C.E., and M. D'Esposito. 2003. Success and failure suppressing reflexive behavior. *Journal of Cognitive Neuroscience* 15 (3): 409–418. <https://doi.org/10.1162/089892903321593126>.
- Davis, K.A., R. Ramachandran, and B.J. May. 2003. Auditory processing of spectral cues for sound localization in the inferior colliculus. *Journal of the Association for Research in Otolaryngology* 4 (2): 148–163. <https://doi.org/10.1007/s10162-002-2002-5>.
- Dehmel, S., Y.L. Cui, and S.E. Shore. 2008. Cross-modal interactions of auditory and somatic inputs in the brainstem and midbrain and their imbalance in tinnitus and deafness. *American Journal of Audiology* 17 (2): S193. [https://doi.org/10.1044/1059-0889\(2008/07-0045\)](https://doi.org/10.1044/1059-0889(2008/07-0045)).
- Deng, Y., I. Choi, B. Shinn-Cunningham, and R. Baumgartner. 2019. Impoverished auditory cues limit engagement of brain networks controlling spatial selective attention. *NeuroImage* 116151. <https://doi.org/10.1016/j.neuroimage.2019.116151>.
- Dietz, M., J.-H. Lestang, P. Majdak, R.M. Stern, T. Marquardt, S.D. Ewert, W.M. Hartmann, and D.F.M. Goodman. 2017. A framework for testing and comparing binaural models. *Hearing Research* 360: 92–106. <https://doi.org/10.1016/j.heares.2017.11.010>.
- Djelani, T., C. Pörschmann, J. Sahrhage, and J. Blauert. 2000. An interactive virtual-environment generator for psychoacoustic research. II: Collection of head-related impulse responses and evaluation of auditory localization. *Acta Acustica United with Acustica* 86 (6): 1046–1053.
- Dokmanić, I., R. Parhizkar, A. Walther, M. Lu, and Y., and Vetterli, M. 2013. Acoustic Echoes Reveal Room Shape. *Proceedings of the National Academy of Sciences of the United States of America* 110 (30): 12186–12191. <https://doi.org/10.1073/pnas.1221464110>.
- Donoho, D.L. 2006. For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics* 59 (6): 797–829.
- Dotov, D.G. 2014. Putting reins on the brain. How the body and environment use it. *Frontiers in Human Neuroscience* 8 (795): 1–12. <https://doi.org/10.3389/fnhum.2014.00795>.
- Epstein, W., and S.J. Rogers (eds.). 1995. *Handbook of Perception and Cognition Perception of Space and Motion*. San Diego: Academic Press.
- Fels, J., J. Oberem, and I. Koch. 2020. Selective binaural attention and attention switching. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 61–90. Cham, Switzerland: Springer and ASA Press.
- Francis, B., and W. Wonham. 1976. The internal model principle of control theory. *Automatica* 12 (5): 457–465. [https://doi.org/10.1016/0005-1098\(76\)90006-6](https://doi.org/10.1016/0005-1098(76)90006-6).
- Friston, K. 2012. Embodied inference and spatial cognition. *Cognitive Processing* 13 (S1): 171–177. <https://doi.org/10.1007/s10339-012-0519-z>.
- Friston, K., J. Kilner, and L. Harrison. 2006. A free energy principle for the brain. *Journal of Physiology* 100 (1–3): 70–87. <https://doi.org/10.1016/j.jphysparis.2006.10.001>.
- Friston, K., T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O'Doherty, and G. Pezzulo. 2016. Active inference and learning. *Neuroscience and Biobehavioral Reviews* 68: 862–879.

- Furuya, H., K. Fujimoto, C. Young Ji, and N. Higa. 2001. Arrival direction of late sound and listener envelopment. *Applied Acoustics* 62 (2): 125–136. [https://doi.org/10.1016/S0003-682X\(00\)00052-9](https://doi.org/10.1016/S0003-682X(00)00052-9).
- Gabriel, K.J., and H.S. Colburn. 1981. Interaural correlation discrimination: I. bandwidth and level dependence. *Journal of the Acoustical Society of America* 69 (5): 1394–1401.
- Gaggioli, A., Bassi, M., and Delle Fave, A. 2003. “Quality of experience in virtual environments”. In *Being There: Concepts, Effects and Measurements of User Presence in Synthetic*, pp. 122–132.
- Genzel, D., U. Firzlaß, L. Wiegrebe, and P.R. MacNeilage. 2016. Dependence of auditory spatial updating on vestibular, proprioceptive, and efference copy signals. *Journal of Neurophysiology* 116 (2): 765–775. <https://doi.org/10.1152/jn.00052.2016>.
- Gordon, C., D.L. Webb, and S. Wolpert. 1992. One cannot hear the shape of a drum. *Bulletin of the American Mathematical Society* 27 (1): 134–138. <https://doi.org/10.1090/S0273-0979-1992-00289-6>.
- Goupell, M.J., and W.M. Hartmann. 2006. Interaural fluctuations and the detection of interaural incoherence: Bandwidth effects. *Journal of the Acoustical Society of America* 119 (6): 3971–3986.
- Gregory, R.L. 1980. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London* 290 (1038): 181–197.
- Griffiths, T.D., and J.D. Warren. 2004. What is an auditory object? *Nature Reviews Neuroscience* 5 (11): 887–892. <https://doi.org/10.1038/nrn1538>.
- Grush, R. 2004. The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences* 27 (3): 377–396.
- Gruters, K.G., and Groh, J.M. 2012. Sounds and beyond: Multisensory and other non-auditory signals in the inferior colliculus. *Front Neural Circuits* 6, 69 <https://doi.org/10.3389/fncir.2012.00096>.
- Hartmann, W.M., and A. Wittenberg. 1996. On the externalization of sound images. *Journal of the Acoustical Society of America* 99 (6): 3678–3688.
- Hartmann, W.M., B. Rakerd, and A. Koller. 2005. Binaural coherence in rooms. *Acta Acustica United with Acustica* 91 (3): 451–462.
- Helmholtz, H. 1954. *On the Sensations of Tone*. NY: Dover Books on Music.
- Henning, G.B. 1974. Detectability of interaural delay in high-frequency complex waveforms. *Journal of the Acoustical Society of America* 55 (1): 84–90.
- Hill, K.T., and L.M. Miller. 2010. Auditory attentional control and selection during cocktail party listening. *Cerebral Cortex* 20 (3): 583–590. <https://doi.org/10.1093/cercor/bhp124>.
- Hinton, G.E., and Z. Ghahramani. 1997. Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society B: Biological Science* 352 (1358): 1177–1190. <https://doi.org/10.1098/rstb.1997.0101>.
- Hládek, u., Tomoriová, B., and Kopčo, N., 2017. Temporal characteristics of contextual effects in sound localization. *Journal of the Acoustical Society of America* 142 (5): 3288–3296. <https://doi.org/10.1121/1.5012746>.
- Hofman, P.M., J.G.A. van Riswick, and A.J. van Opstal. 1998. Relearning sound localization with new ears. *Nature Neuroscience* 1 (5): 417–421. <http://dx.doi.org/10.1038/1633>.
- Katz, B.F.G. 2001. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *Journal of the Acoustical Society of America* 110 (5): 2440–2448. <https://doi.org/10.1121/1.1412440>.
- Kayser, C., and N.K. Logothetis. 2007. Do early sensory cortices integrate cross-modal information? *Brain Structure and Function* 212 (2): 121–132. <https://doi.org/10.1007/s00429-007-0154-0>.
- Keating, P., J.C. Dahmen, and A.J. King. 2015. Complementary adaptive processes contribute to the developmental plasticity of spatial hearing. *Nature Neuroscience* 18 (2): 185–187. <https://doi.org/10.1038/nn.3914>.
- Klier, E.M. 2003. Three-dimensional eye-head coordination is implemented downstream from the superior colliculus. *Journal of Neurophysiology* 89 (5): 2839–2853. <https://doi.org/10.1152/jn.00763.2002>.

- Klockgether, S., and S. van de Par. 2014. A model for the prediction of room acoustical perception based on the just noticeable differences of spatial perception. *Acta Acustica united with Acustica* 100: 964–971. <https://doi.org/10.3813/AAA.918776>.
- Knudsen, E.I. 2007. Fundamental components of attention. *Annual Review of Neuroscience* 30 (1): 57–78. <https://doi.org/10.1146/annurev.neuro.30.051606.094256>.
- Koffka, K. 1935. *Principles of Gestalt psychology*. London: Mimesis Int.
- Kolarik, A.J., B.C.J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan. 2016. Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, and Psychophysics* 78 (2): 373–395. <https://doi.org/10.3758/s13414-015-1015-1>.
- Kondo, H.M., D. Pressnitzer, I. Toshima, and M. Kashino. 2012. Effects of self-motion on auditory scene analysis. *Proceedings of the National Academy of Sciences of the United States of America* 109 (17): 6775–6780. <https://doi.org/10.1073/pnas.1112852109>.
- Kreuzer, W., P. Majdak, and Z. Chen. 2009. Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range. *Journal of the Acoustical Society of America* 126 (3): 1280–1290. <https://doi.org/10.1121/1.3177264>.
- Kuhl, W. 1978. Räumlichkeit als Komponente des Raumeindrucks (Spaciousness (spatial impression) as a component of total room impression). *Acustica* 40 (3): 167–181.
- Kuhn, G.F. 1977. Model for the interaural time differences in the azimuthal plane. *Journal of the Acoustical Society of America* 62 (1): 157–167. <https://doi.org/10.1121/1.381498>.
- Kulkarni, A., S.K. Isabelle, and H.S. Colburn. 1999. Sensitivity of human subjects to head-related transfer-function phase spectra. *Journal of the Acoustical Society of America* 105 (5): 2821–2840.
- Laback, B., M. Dietz, and P. Joris. 2017. Temporal effects in interaural and sequential level difference perception. *Journal of the Acoustical Society of America* 142 (5): 3267–3283. <https://doi.org/10.1121/1.5009563>.
- Lavandier, M., S. Jelfs, J.F. Culling, A.J. Watkins, A.P. Raimond, and S.J. Makin. 2012. Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. *Journal of the Acoustical Society of America* 131 (1): 218–231. <https://doi.org/10.1121/1.3662075>.
- Leaver, A.M., J. Van Lare, B. Zielinski, A.R. Halpern, and J.P. Rauschecker. 2009. Brain activation during anticipation of sound sequences. *The Journal of Neuroscience* 29 (8): 2477–2485. <https://doi.org/10.1523/JNEUROSCI.4921-08.2009>.
- Lee, C.C. 2015. Exploring functions for the non-lemniscal auditory thalamus. *Front Neural Circuits* 9. <https://doi.org/10.3389/fncir.2015.00069>.
- Lee, C.C., and S.M. Sherman. 2010. Topography and physiology of ascending streams in the auditory tectothalamic pathway. *Proceedings of the National Academy of Sciences of the United States of America* 107 (1): 372–377. <https://doi.org/10.1073/pnas.0907873107>.
- Leung, J., Wei, V., Burgess, M., and Carlile, S. 2016. Head tracking of auditory, visual, and audio-visual targets. *Frontiers in Neuroscience* 9: 493. <https://doi.org/10.3389/fnins.2015.00493>.
- Lokki, T., and J. Pätynen. 2020. Auditory spatial impression in concert halls. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 173–202. Cham, Switzerland: Springer and ASA Press.
- Macpherson, E.A., and J.C. Middlebrooks. 2002. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *Journal of the Acoustical Society of America* 111 (5): 2219–2236. <https://doi.org/10.1121/1.1471898>.
- Macpherson, E.A., and J.C. Middlebrooks. 2003. Vertical-plane sound localization probed with ripple-spectrum noise. *Journal of the Acoustical Society of America* 114 (1): 430–445. <https://doi.org/10.1121/1.1582174>.
- Macpherson, E.A., and A.T. Sabin. 2007. Binaural weighting of monaural spectral cues for sound localization. *Journal of the Acoustical Society of America* 121 (6): 3677–3688. <https://doi.org/10.1121/1.2722048>.
- Majdak, P., P. Balazs, and B. Laback. 2007. Multiple exponential sweep method for fast measurement of head-related transfer functions. *Journal of the Audio Engineering Society* 55: 623–637.

- Majdak, P., M.J. Goupell, and B. Laback. 2010. 3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Attention, Perception, and Psychophysics* 72 (2): 454–69. <https://doi.org/10.3758/APP.72.2.454>.
- Majdak, P., T. Carpentier, R. Nicol, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorff, H. Ziegelwanger, and M. Noisternig. 2013a. Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions. In *Proceeding of 134th Convention Audio Engineering Society*, 8880. Roma: Italy.
- Majdak, P., T. Walder, and B. Laback. 2013b. Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions. *Journal of the Acoustical Society of America* 134 (3): 2148–2159. <https://doi.org/10.1121/1.4816543>.
- Majdak, P., R. Baumgartner, and B. Laback. 2014. Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. *Frontiers in Psychology* 5 (319): 1–10. <https://doi.org/10.3389/fpsyg.2014.00319>.
- Mason, R., T. Brookes, and F. Rumsey. 2005. Frequency dependency of the relationship between perceived auditory source width and the interaural cross-correlation coefficient for time-invariant stimuli. *Journal of the Acoustical Society of America* 117 (3): 1337–1350. <https://doi.org/10.1121/1.1853113>.
- May, B.J. 2000. Role of the dorsal cochlear nucleus in the sound localization behavior of cats. *Hearing Research* 148 (1–2): 74–87.
- McAnally, K.I., and R.L. Martin. 2014. Sound localization with head movement: Implications for 3-d audio displays. *Frontiers in Neuroscience* 8 (210): 1–6. <https://doi.org/10.3389/fnins.2014.00210>.
- Meloni, E.G., and M. Davis. 1998. The dorsal cochlear nucleus contributes to a high intensity component of the acoustic startle reflex in rats. *Hearing Research* 119 (1–2): 69–80.
- Mendonça, C. 2014. A review on auditory space adaptations to altered head-related cues. *Auditory Cognitive Neuroscience* 8: 219. <https://doi.org/10.3389/fnins.2014.00219>.
- Mendonça, C. 2020. Psychophysical models of sound localisation with audiovisual interactions. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 289–314. Cham, Switzerland: Springer and ASA Press.
- Mendonça, C., G. Campos, P. Dias, and J.A. Santos. 2013. Learning auditory space: Generalization and long-term effects. *PLoS One* 8 (10): e77900. <https://doi.org/10.1371/journal.pone.0077900>.
- Micheyl, C., R.P. Carlyon, A. Gutschalk, J.R. Melcher, A.J. Oxenham, J.P. Rauschecker, B. Tian, and E. Courtenay Wilson. 2007. The role of auditory cortex in the formation of auditory streams. *Hearing Research* 229 (1–2): 116–131. <https://doi.org/10.1016/j.heares.2007.01.007>.
- Middlebrooks, J.C. (2009). Auditory system: Central pathways. In *Encyclopedia of Neuroscience* Oxford: Academic Press, pp. 745–752.
- Middlebrooks, J.C. 2015. Sound localization. *Handbook of Clinical Neurology* 129: 99–116. <https://doi.org/10.1016/B978-0-444-62630-1.00006-8>.
- Miller, L.M., and G.H. Recanzone. 2009. Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. *Proceedings of the National Academy of Sciences of the United States of America* 106 (14): 5931–5935. <https://doi.org/10.1073/pnas.0901023106>.
- Möller, S., and Raake, A. 2014. *Quality of Experience: Advanced Concepts, Applications and Methods*. Springer.
- Møller, H., M.F. Sørensen, D. Hammershøi, and C.B. Jensen. 1995. Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society* 43: 300–321.
- Moore, A.H., Brookes, M., and Naylor, P.A. 2013. Room geometry estimation from a single channel acoustic impulse response. In *Proceedings of European Signal Processing Conference EUSIPCO*, 1–5.
- Morimoto, M., K. Iida, and K. Sakagami. 2001. The role of reflections from behind the listener in spatial impression. *Applied Acoustics* 62 (2): 109–124. [https://doi.org/10.1016/S0003-682X\(00\)00051-7](https://doi.org/10.1016/S0003-682X(00)00051-7).

- Muniak, M.A., and D.K. Ryugo. 2014. Tonotopic organization of vertical cells in the dorsal cochlear nucleus of the CBA/J mouse: Tonotopic organization of vertical cells in the DCN. *The Journal of Comparative Neurology* 522 (4): 937–949. <https://doi.org/10.1002/cne.23454>.
- Nelken, I., J. Bizley, S.A. Shamma, and X. Wang. 2014. Auditory cortical processing in real-world listening: The auditory system going real. *The Journal of Neuroscience* 34 (46): 15135–15138. <https://doi.org/10.1523/JNEUROSCI.2989-14.2014>.
- Noble, W., and S. Gatehouse. 2006. Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the Speech, Spatial, and Qualities of Hearing scale (SSQ). *International Journal of Audiology* 45 (3): 172–181. <https://doi.org/10.1080/14992020500376933>.
- Oberfeld, D., and Klöckner-Nowotny, F. 2016. Individual differences in selective attention predict speech identification at a cocktail party. *eLife* 5: 16747. <https://doi.org/10.7554/eLife.16747>.
- Okano, T., L.L. Beranek, and T. Hidaka. 1998. Relations among interaural cross-correlation coefficient (IACCE), lateral fraction (LFE), and apparent source width (ASW) in concert halls. *Journal of the Acoustical Society of America* 104 (1): 255–265. <https://doi.org/10.1121/1.423955>.
- Pätynen, J., and T. Lokki. 2016. Concert halls with strong and lateral sound increase the emotional impact of orchestra music. *Journal of the Acoustical Society of America* 139 (3): 1214–1224. <https://doi.org/10.1121/1.4944038>.
- Peck, C.K. 1996. Visual-auditory integration in cat superior colliculus: Implications for neuronal control of the orienting response. *Progress in Brain Research* 112: 167–177.
- Pecka, M., C. Leibold, and B. Grothe. 2020. Biological aspects of perceptual space formation. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 151–172. Cham, Switzerland: Springer and ASA Press.
- Perrett, S., and W. Noble. 1995. Available response choices affect localization of sound. *Perception and Psychophysics* 57: 150–158.
- Perrott, D.R., and K. Saberi. 1990. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *Journal of the Acoustical Society of America* 87 (4): 1728–1731. <https://doi.org/10.1121/1.399421>.
- Rauschecker, J.P., and Tian, B. 2000. Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America* 97 (22): 11800–11806. <https://doi.org/10.1073/pnas.97.22.11800>.
- Rauschecker, J.P. 2011. An expanded role for the dorsal auditory pathway in sensorimotor control and integration. *Hearing Research* 271 (1–2): 16–25. <https://doi.org/10.1016/j.heares.2010.09.001>.
- Reichinger, A., Majdak, P., Sablatnig, R., and Maierhofer, S. 2013. Evaluation of methods for optical 3-D scanning of human pinnas. *International Conference on 3D Vision*, 390–397. <https://doi.org/10.1109/3DV.2013.58>.
- Ruggles, D., H. Bharadwaj, and B.G. Shinn-Cunningham. 2012. Why middle-aged listeners have trouble hearing in everyday settings. *Current Biology* 22 (15): 1417–1422. <https://doi.org/10.1016/j.cub.2012.05.025>.
- Salminen, N.H., M. Takanen, O. Santala, J. Lamminsalo, A. Altoè, and V. Pulkki. 2015. Integrated processing of spatial cues in human auditory cortex. *Hearing Research* 327: 143–152. <https://doi.org/10.1016/j.heares.2015.06.006>.
- Schechtman, E., T. Shrem, and L.Y. Deouell. 2012. Spatial localization of auditory stimuli in human auditory cortex is based on both head-independent and head-centered coordinate systems. *The Journal of Neuroscience* 32 (39): 13501–13509. <https://doi.org/10.1523/JNEUROSCI.1315-12.2012>.
- Schultz, D.P., and S.E. Schultz. 2015. *A History of Modern Psychology*, 11th ed. Boston, MA: Cengage Learning.
- Shinn-Cunningham, B.G., S. Santarelli, and N. Kopco. 2000. Tori of confusion: Binaural localization cues for sources within reach of a listener. *Journal of the Acoustical Society of America* 107 (3): 1627–1636. <http://view.ncbi.nlm.nih.gov/pubmed/10738816>.
- Shinn-Cunningham, B.G. 2008. Object-based auditory and visual attention. *Trends in Cognitive Sciences* 12 (5): 182–186. <https://doi.org/10.1016/j.tics.2008.02.003>.

- Shore, S.E. 2005. Multisensory integration in the dorsal cochlear nucleus: Unit responses to acoustic and trigeminal ganglion stimulation. *European Journal of Neuroscience* 21 (12): 3334–3348. <https://doi.org/10.1111/j.1460-9568.2005.04142.x>.
- Singla, S., C. Dempsey, R. Warren, A.G. Enikolopov, and N.B. Sawtell. 2017. A cerebellum-like circuit in the auditory system cancels responses to self-generated sounds. *Nature Neuroscience* 20 (7): 943–950. <https://doi.org/10.1038/nn.4567>.
- Skottun, B.C., T.M. Shackleton, R.H. Arnott, and A.R. Palmer. 2001. The ability of inferior colliculus neurons to signal differences in interaural delay. *Proceedings of the National Academy of Sciences of the United States of America* 98 (24): 14050–14054. <https://doi.org/10.1073/pnas.241513998>.
- Slama, M.C.C., and B. Delgutte. 2015. Neural coding of sound envelope in reverberant environments. *The Journal of Neuroscience* 35 (10): 4452–4468. <https://doi.org/10.1523/JNEUROSCI.3615-14.2015>.
- Slater, M. 2003. A note on presence terminology. *Presence Connect* 3.
- Slee, S.J., and E.D. Young. 2013. Linear processing of interaural level difference underlies spatial tuning in the nucleus of the brachium of the inferior colliculus. *The Journal of Neuroscience* 33 (9): 3891–3904. <https://doi.org/10.1523/JNEUROSCI.3437-12.2013>.
- Slee, S.J., and E.D. Young. 2014. Alignment of sound localization cues in the nucleus of the brachium of the inferior colliculus. *Journal of Neurophysiology* 111 (12): 2624–2633. <https://doi.org/10.1152/jn.00885.2013>.
- Snyder, J.S., and M. Elhilali. 2017. Recent advances in exploring the neural underpinnings of auditory scene perception. *Annals of the New York Academy of Sciences* 1396 (1): 39–55. <https://doi.org/10.1111/nyas.13317>.
- Sokolov, E. 2001. Orienting response. In *International Encyclopedia of the Social & Behavioral Sciences*. Pergamon: Elsevier, 10978–10981.
- Strack, F., and R. Deutsch. 2004. Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review* 8 (3): 220–247.
- Straka, M.M., S. Schmitz, and H.H. Lim. 2014. Response features across the auditory midbrain reveal an organization consistent with a dual lemniscal pathway. *Journal of Neurophysiology* 112 (4): 981–998. <https://doi.org/10.1152/jn.00008.2014>.
- Strutt alias Lord Rayleigh, J.W. 1876. Our perception of the direction of a source of sound. *Proc Musical Association* 2: 75–84.
- Sutojo, S., J. Thiemann, A. Kohlrausch, and S. van de Par. 2020. Auditory gestalt rules and their application. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 33–59. Cham, Switzerland: Springer and ASA Press.
- Szabó, B.T., S.L. Denham, and I. Winkler. 2016. Computational models of auditory scene analysis: A Review. *Frontiers in Neuroscience* 10. <https://doi.org/10.3389/fnins.2016.00524>.
- Trapeau, R., and M. Schönwiesner. 2015. Adaptation to shifted interaural time differences changes encoding of sound location in human auditory cortex. *NeuroImage* 118: 26–38. <https://doi.org/10.1016/j.neuroimage.2015.06.006>.
- Trapeau, R., V. Aubrais, and M. Schönwiesner. 2016. Fast and persistent adaptation to new spectral cues for sound localization suggests a many-to-one mapping mechanism. *Journal of the Acoustical Society of America* 140 (2): 879–890. <https://doi.org/10.1121/1.4960568>.
- van der Heijden et al. 2019. <https://www.nature.com/articles/s41583-019-0206-5#article-info>.
- Viaud-Delmon, I., and Warusfel, O. 2014. From ear to body: The auditory-motor loop in spatial cognition. *Auditory Cognitive Neuroscience* 8: 283. <https://doi.org/10.3389/fnins.2014.00283>.
- Vorländer, M., and Shinn-Cunningham, B. 2014. Virtual auditory displays. In *Handbook of Virtual Environment Technology*, ed. Hale, K.S., and K.M. Stanney, 2nd ed., 87–114. Boca Raton: CRC Press.
- Wang, D., and Brown, G.J. 2006. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.

- Wenzel, E.M., Fisher, S.S., Stone, P.K., and Foster, S.H. 1990. A system for three-dimensional acoustic visualization in a virtual environment workstation. In *Proceedings of the 1st Conference on Visualization'90*. IEEE Computer Society Press, 329–337.
- Whitmer, W.M., B.U. Seeber, and M.A. Akeroyd. 2013. Measuring the apparent width of auditory sources in normal and impaired hearing. *Advances in Experimental Medicine and Biology* 787: 303–310. https://doi.org/10.1007/978-1-4614-1590-9_34.
- Winkler, I., S.L. Denham, and I. Nelken. 2009. Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences* 13 (12): 532–540. <https://doi.org/10.1016/j.tics.2009.09.003>.
- Witmer, B.G., and M.J. Singer. 1998. Measuring presence in virtual environments: A presence questionnaire. *Presence* 7 (3): 225–240.
- Woods, J.W. 1964. Behavior of chronic decerebrate rats. *Journal of Neurophysiology* 27: 635–644. <https://doi.org/10.1152/jn.1964.27.4.635>.
- Xie, B. 2013. *Head-Related Transfer Function and Virtual Auditory Display*. Plantatation, FL: J. Ross Publishing.
- Yao, J.D., P. Bremen, and J.C. Middlebrooks. 2015. Transformation of spatial sensitivity along the ascending auditory pathway. *Journal of Neurophysiology* 113 (9): 3098–3111. <https://doi.org/10.1152/jn.01029.2014>.
- Yost, W.A. 1974. Discriminations of interaural phase differences. *Journal of the Acoustical Society of America* 55 (6): 1299–1303. <https://doi.org/10.1121/1.1914701>.
- Yost, W.A., X. Zhong, and A. Najam. 2015. Judging sound rotation when listeners and sounds rotate: Sound source localization is a multisystem process. *Journal of the Acoustical Society of America* 138 (5): 3293–3310. <https://doi.org/10.1121/1.4935091>.
- Ziegelwanger, H., and P. Majdak. 2014. Modeling the direction-continuous time-of-arrival in head-related transfer functions. *Journal of the Acoustical Society of America* 135 (3): 1278–1293. <https://doi.org/10.1121/1.4863196>.
- Ziegelwanger, H., P. Majdak, and W. Kreuzer. 2015. Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization. *Journal of the Acoustical Society of America* 138 (1): 208–222. <https://doi.org/10.1121/1.4922518>.
- Ziegelwanger, H., W. Kreuzer, and P. Majdak. 2016. A priori mesh grading for the numerical calculation of the head-related transfer functions. *Applied Acoustics* 114: 99–110. <https://doi.org/10.1016/j.apacoust.2016.07.005>.

Biological Aspects of Perceptual Space Formation



Michael Pecka, Christian Leibold and Benedikt Grothe

Abstract Traditional ideas of how auditory space is formed and represented in the brain have been dominated by the concept of topographically arranged neuronal maps—similar to what is known from the visual system. Specifically, it had canonically been assumed that the brain’s representation of the location of sound sources is “hard-wired”, that is a specific location in space relative to the head is encoded by a particular sub-set of neurons tuned to that head angle. However, recent experimental findings strongly contradict this assumption for the computation of sound location in mammals (including humans). These data rather suggest a “relative” spatial code that favors the determination of changes in location over its absolute position. Here we explain the mechanisms underlying neuronal spatial sensitivity in mammals and summarize the data that led to this paradigm shift. We further explain that a consideration of evolutionary constraints of spatial cue use and their processing strategies is crucial for the understanding of the concepts underlying auditory spatial representation in mammals. Finally, we review recent neurophysiological and psychophysical findings demonstrating pronounced context-dependent plasticity in the neuronal coding and perception. We conclude that mammalian spatial hearing is based on a relative representation of auditory space, which has significant implications for how we localize sound sources in complex environments.

1 Introduction

The human nervous system allows the perception of space via multiple modalities including somato-sensation, vestibular inputs, motor efference copies, proprioception, vision and audition. The latter two stand out in that they allow us to also perceive

M. Pecka (✉) · C. Leibold · B. Grothe
Division of Neurobiology, Department Biology II,
Ludwig-Maximilians-Universität München, Großhaderner Str. 2-4,
82152 Martinsried-Planegg, Germany
e-mail: pecka@biologie.uni-muenchen.de

C. Leibold
Bernstein Center for Computational Neuroscience Munich,
Großhaderner Str. 2-4, 82152 Martinsried, Germany

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_6

distant components of space. Yet spatial processing is entirely dissimilar between the visual and the auditory domain: the retina innately provides information about spatial relationships between the sensory source(s), as photoreceptors are arranged on a 2-dimensional surface such that neighboring receptors encode neighboring stimulus positions. In audition, the cochlea does not provide such an inherent space representation. In fact, neighboring cochlear locations encode adjacent sound frequencies, and consequently, auditory space has to be computed by dedicated circuits downstream of the receptor organ in the brain by exploiting spectral and temporal sound features.

2 Creating a Sense of Auditory Space

2.1 Cues for Spatial Hearing and Their Neural Processing

Sound localization in mammals is based on two complementary yet distinct neuronal computations of analyzing the acoustic waveform (Fig. 1). The first constitutes a spectral analysis in which the comparison of sound energy across different frequency bands arriving at each ear provides for sound localization abilities in the vertical dimension and distinctions between sources in the front and rear. Although better performance based on frequency-spectra may be possible using both ears, it represents an essentially monaural cue for sound localization, generated largely by the direction-specific attenuation of particular frequencies by the pinna and concha of the outer ear. The second neuronal mechanism for sound localization is based on detecting and comparing differences in the signals between the two ears (or more precisely, between the two cochleae). This binaural computation, which takes place mainly within narrowband sound-frequency channels, underlies sound localization in the horizontal plane, i.e. allows for determination of the lateral angle. Two interaural differences are available to such binaural analysis. We will next introduce the neuronal pathways and mechanism underlying these monaural and binaural means of sound localization, as the remarkably high plasticity that characterizes mammalian spatial hearing can be directly related to these mechanisms. It should be noted that for reasons of comprehensibility, we here focus on mammalian sound localization mechanisms and coding strategies only. For a comparative analysis of the mammalian and avian system, we refer the reader to Grothe and Pecka (2014).

Monaural Sound Localization

Neurons in the dorsal division of the Cochlear Nucleus (DCN) are particularly specialized for processing spectral cues. In this context, more recent investigations have begun to examine neural coding of spectral cues for localization in the midbrain center of the auditory pathway, the Inferior Colliculus (IC). A particular focus of many investigations has been put on how coding of spectral cues is modified between the lower brainstem and the IC (Fig. 1a2–a3). Responses of the so-called *Type-IV* neurons of the DCN appear to be determined by a dedicated neural circuit within the DCN itself (Oertel and Young 2004) (Fig. 1a2). When considering the response rates of these cell type as a function of sound intensity and frequency, type-IV neurons show

a small “island” (i.e. combination of frequency and intensity) of near-threshold activation around their characteristic frequency (CF; the frequency at which the threshold for tone-evoked responses is lowest), with a prominent inhibitory input at higher sound intensities (Imig et al. 2000; Davis et al. 2003). This inhibition is more broadly tuned than the excitatory response area of Type-IV neurons through convergent input from multiple (differently-tuned) *Type-II* DCN neurons. The convergence of excitatory inputs from primary auditory nerve fibres (ANFs) and inhibition derived from *Type-II* DCN renders Type-IV neurons particularly sensitive to notches in the acoustic spectrum, presumably those generated by the interaction of sound with the head and pinna (Imig et al. 2000; Young et al. 1992). How might neurons encode the potential cues for sound source elevation that these notches provide? The answer lies in understanding the output neurons to which Type-IV neurons project (Fig. 1a3). Neurons in the IC described as *Type-O*—they possess a circumscribed frequency-vs-intensity response area—are the main target of DCN Type-IV neurons (Oertel and Young 2004). When stimulated with pure tones, Type-O neurons, like Type-IV neurons, show a largely inhibitory receptive field with a small “island” of excitation at low stimulus intensities. However, when stimulated with broadband sounds containing a spectral notch, mimicking the effect of the direction-specific head-related transfer function (HRTF), they respond with essentially the opposite characteristics as compared to Type-IV neurons in the DCN. Actually, they show considerable excitatory responses for a specific single-notch frequency, particularly at higher sound intensities, flanked by inhibitory regions generated by all other notch frequencies (Oertel and Young 2004). Thus, IC neurons appear to show an essentially unambiguous response to the frequency of a spectral notch, and the pathway from Type IV neurons in the DCN to Type-O neurons in the IC seems to be uniquely specialized for processing directionally-dependent spectral features generated by the HRTF.

Interestingly, experiments in humans and rodents have revealed remarkable perceptual and neuronal plasticity in using spectral cues for localization even in the horizontal plane (for review see (Mendonça 2014)). Monaural deprivation assays in adult human listeners demonstrated that, given extensive training, subjects can learn to use monaural cues to achieve respectable sound-localization performance (Keating et al. 2016), and even to (re)interpret altered interaural cues—see below. These remarkable findings mimic developmental mechanisms found in ferrets (Keating and King 2013; Keating et al. 2013; Keating and King 2015) and thus highlight the adaptive nature of the spatial code in mammals. They also help explain why human sound localization ability apparently tolerates acute (e.g., when wearing a hat) and chronic (age-dependent changes in ear shape) alterations in the HRTFs.

The following discussion elaborates on the importance of context-dependent adaption of spatial coding (Mendonça 2014; Keating and King 2015) and how it constitutes a dedicated mechanism for the processing of binaural cues.

Binaural Spatial Sensitivity via Coincidence Detection

The fundamental motive of binaural spatial processing is the neuronal comparison of specific physical sound properties between the left and right ear by dedicated populations in the auditory brainstem. Specifically, individual neurons in the lateral and

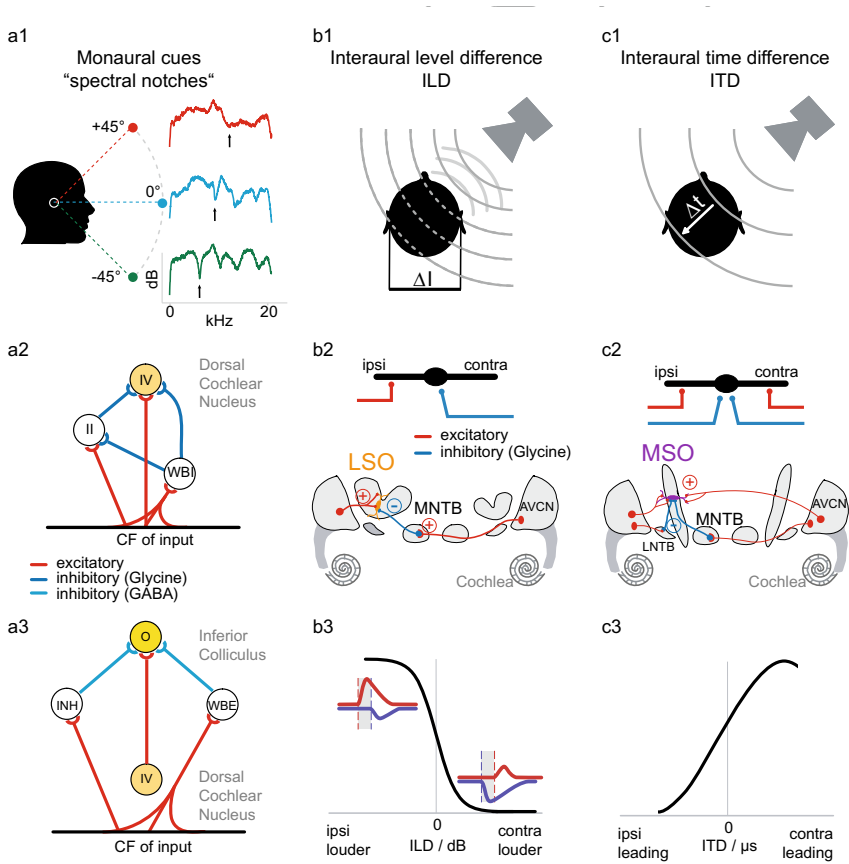


Fig. 1 Cues and neuronal circuits for sound localization. **a** Spectral notches in broadband stimuli can be used for localization in the vertical plane. **a1** The notches are detected by complex microcircuits in the dorsal Cochlear Nucleus, **a2**, and Inferior Colliculus, **a3**, by integrating excitatory and inhibitory inputs. **b, c** For localization in the horizontal plane, binaural cues are required. **b1** Interaural level differences (ILD) are generated by the partial reflection and absorption of sound waves by the head. Substantial ILDs are generated for frequencies approx. >2 kHz in humans. **b2** ILDs are computed by neurons in the Lateral Superior Olive (LSO) by comparison of excitatory and inhibitory inputs from the two ears. **b3** Schematic ILD-response function of LSO neurons. The response rate is determined both by the relative amplitude and timing of the two inputs. **c1** For low frequencies (<2 kHz), interaural time differences (ITDs), i.e. differences in the arrival time of sound waves at the two ears in the range of ten to hundreds of microseconds, are used for sound localization. **c2** ITDs are detected by neurons in the Medial Superior Olive (MSO) by coincidence detection of excitatory and inhibitory inputs from both ears. **c3** Similar to ILDs in the LSO, the coding of ITDs in the MSO is panoramic, that is, spans the entire range of physiological ITDs

the medial superior olive (LSO and MSO, respectively) assess the relative amplitude and coincidence between inputs from the two ears (Fig. 1b1–b3, c1–c3). The binaural computation in both LSO and MSO is based on precise interactions of glutamatergic excitation and glycinergic inhibition, and consequently the two circuits share the ipsilateral excitatory as well as the prominent contralateral inhibitory pathway. MSO and LSO are thought to analyze distinct binaural cues related to the frequency of the incoming sound. That is, differences in the relative level of sounds between the ears—interaural level differences (ILDs)—are processed in the LSO. The magnitude of ILDs as a function of head angle is frequency dependent, where larger ILDs are produced at higher frequencies. For low frequencies (approx. <2 kHz), a second parameter is predominately used, namely the head-angle specific microsecond differences in the time-of-arrival—interaural time difference (ITD)—of sounds at the two ears. Processing of ITDs requires higher input-timing accuracy in the range of only tens of microseconds and is mainly performed in the MSO.

Evolutionary Aspects Determine the Coding Strategy

To understand the nature of mammalian auditory-space representation, it is important to appreciate that the evolutionary starting point for spatial hearing and neuronal processing in mammals was their anatomy/morphology, which dictated the use of one particular interaural cue and, in turn, determined the processing and coding design for mammalian binaural hearing.

The ancestors of mammals in the Late Triassic were animals smaller than laboratory mice (Allin 1975; Clack 1997). Interestingly, during this phase the originally much larger middle-ear bones shrank isometrically with the rest of the skull to a size suitable for transmitting sounds, and they have allometrically remained in this state despite the ensuing changes in overall body size (Hylander and Crompton 1986). For reasons of size-related resonance of the middle ear bones (Rosowski 1991), it follows that early mammals were high-frequency hearing animals. This *specialization* of mammals to hear high frequencies tended to increase rather than diminish during evolution. In fact, the audiograms of recent mammals of various groups indicates that their hearing range almost exclusively extended into the high-frequency range (Grothe and Pecka 2014). Notably, such extension of hearing range to ever higher frequencies significantly improves the ability to use spectral cues for localization in the vertical plane (given a co-evolution of asymmetric pinnae—see paragraph above). Since localization in the vertical axis is of utmost importance for small prey animals, reliable HRTF-based localization may well have been a crucial evolutionary pressure on the hearing range of small early mammals. The second advantage of mainly high-frequency hearing is that even the smallest mammals have always experienced significant ILDs (Erulkar 1972; Harnischfeger et al. 1985). On the other hand, their tiny heads produced ITDs of a few tens of microseconds at best. Even today, most small mammals rely almost entirely on ILDs. The neuronal structure responsible for the initial processing of ILDs, the LSO, is homogenous in all terrestrial mammals investigated (Tollin 2003; Grothe et al. 2010). In contrast, the MSO exhibits significant differences in shape and size, which are likely to be related to the hearing range in the respective species—low- versus high-frequency sensitivity (Grothe 2000). Significant selection pressure to use ITDs existed only relatively

late during the evolution of mammals, probably in relation to increasing body size, which not only conditioned production of low-frequency communication calls but also necessitated larger territories. Thus, it is of advantage that low-frequency sounds travel long distances.

Early mammals, however, could most probably hear high-frequency sounds and had relatively small heads. Hence, ILDs were the only binaural cues available to them for azimuthal sound localization. This suggests that the ancestral neuronal structure used to process interaural spatial information was devoted to ILD detection. As mentioned earlier, ILD sensitivity is generated by the LSO in the brainstem, whose bipolar neurons are the initial site of binaural convergence (Fig. 1b2) (Galambos et al. 1959; Boudreau and Tsuchitani 1968; Tsuchitani and Boudreau 1969). They integrate excitatory (glutamatergic) inputs from the ipsilateral antero-ventral cochlear nucleus (AVCN) with inhibitory (glycinergic) inputs coming from the ipsilateral medial nucleus of the trapezoid body (MNTB) via highly myelinated and rapidly conducting axons and the giant calyx of Held synapses. The MNTB itself is innervated by the contralateral cochlear nucleus. Accordingly, LSO response rates (i.e. the number of action potentials elicited per unit time) are highest for ipsilateral sound source locations that create positive ILDs. In other words, a high sound level at the ipsilateral ear allows the excitatory pathway to be fully activated, whereas the sound level at the farther ear is greatly attenuated by the skull, and thus activation of the contralateral inhibitory pathway is comparatively much smaller—Fig. 1b3. More importantly, response rates are modulated as a function of the ILD. Most LSO neurons are completely inhibited from spiking at ILDs favoring the contralateral ear (negative ILDs), when the sound intensity at the inhibitory ear is highest and lowest at the excitatory ear. For any ILDs in between, response functions typically take the shape of a sigmoid, generating high sensitivity for small changes in ILD along the slope of the function (Tollin 2003).

Crucially, beyond the gauging of input strength, the comparison of the relative timing of the excitatory and inhibitory inputs in the LSO is an integral part of ILD computation (Tollin 2003; Ashida et al. 2016). Because the latency of the input to the auditory brainstem depends on the sound amplitude at the respective ear, any changes in the ILD (i.e. the sound source position) consequently entail a change in the relative arrival times of the respective inputs at the LSO (Park et al. 1996)—Fig. 1b3. Thus, ILD processing by LSO neurons constitutes of the gauging of both the input levels and the input timing. Accordingly, LSO neurons are also sensitive to ITDs (Finlayson and Caspary 1991; Tollin and Yin 2005). This suggests that the (ancestral) circuit of the LSO for ILD processing was to some extent pre-adapted to also be sensitive for microsecond-precise temporal (i.e. ITD) processing. Consequently, when mammals—particularly their heads and thus the inter-ear-distance—increased during evolution, and larger ITDs were experienced, this pre-adaptation might have served beneficial to the development of the MSO circuit that is dedicated to ITD processing (Grothe and Pecka 2014)—Fig. 2. Accordingly, the MSO circuit still shares the two prominent inputs with the LSO circuit, namely, ipsilateral AVCN input and contralateral driven inhibition via the MNTB, but additionally incorporates ipsilateral inhibition from the lateral nucleus of the trapezoid body as well as contralateral excitation—Fig. 1c2.

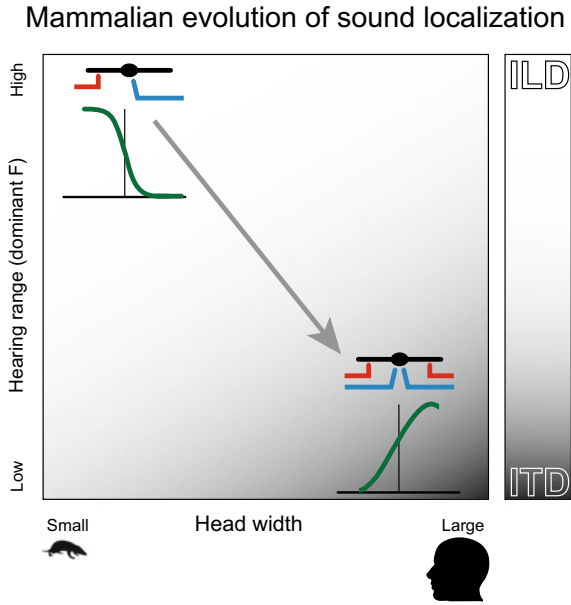


Fig. 2 Mammalian evolution of binaural spatial processing is dictated by cue accessibility. Head width (**abscissa**) and hearing range (**ordinate**) during the time of the middle-ear development define the interaural cue that is most easily exploitable for horizontal sound localization by mammals—see the scale: **white...ILD, gray...ITD**). The interaural cues, in turn, shaped the emergence of a distinct neuronal mechanism optimized for the processing and encoding of the particular cues. Early mammals were very small and had a high-frequency hearing system. Therefore, they used ILDs as the original interaural cue. Subsequent evolutionary changes in the head size and/or the audible frequency range—such as in humans—allowed the use of ITDs. However, the neuronal mechanisms underlying precise temporal integration of excitatory and inhibitory inputs remained similar to early high-frequency-hearing mammals (as schematized by the **black neuron** receiving **red** and **blue inputs**, respectively). Likewise, the coding principles are the same, that is, a hemispheric population code (schematized by idealized tuning functions)—compare Fig. 1 for details

Biophysical Aspects of Neuronal Coincidence Detection

Since the MSO (in contrast to the LSO) receives bilateral excitation, it implements a neural coincidence detection mechanism that on its own is already ITD sensitive (Jeffress 1948). Excitatory inputs that arrive coincident in time evoke maximal firing rates, and the firing rates decrease for increasing temporal disparities between the two inputs. The ITD resolution that such a coincidence mechanism can achieve, depends on, (1), the biophysical filtering properties of the neuronal elements (synaptic transmission and membrane integration) and, (2), the temporal structure of the synaptic input. The biophysical properties of MSO neurons indeed make them remarkably fast: Membrane voltage integration occurs with a leakage time constant of few 100 μ s (for comparison: LSO neurons have time constants of about 1 ms; Fischer et al. (2018)), and synaptic transmission of excitatory synapses is on the same order of magnitude (Scott et al. 2005; Couchman et al. 2010). Conversely, the

time structure of the input is determined by the cochlear frequency channel, therefore lower frequencies necessarily produce broader peaks in the ITD tuning functions. For single neurons, Fisher-type information is largest at the slope (Harper and McAlpine 2004) and therefore it is optimal to position the slopes of the tuning curves near midline, i.e., shift the peaks outside the physiological range of ITDs (which is in first approximation given by the inter-ear distance). Indeed, such alignment of peaks and slopes is experimentally observed (see below), yet the mechanisms underlying these peak shifts in particular the role of the prominent inhibitory inputs are highly debated (Brand et al. 2002; Joris and Yin 2007; Pecka et al. 2008; van der Heijden et al. 2013; Franken et al. 2014; Myoga et al. 2014).

Importantly, since it is derived from LSO processing, the spatial code for ITDs that is generated in the MSO mirrors that of the LSO, i.e. exhibits broad, linear modulation of spike rates across the physiological range of ITDs, while peak response rates are typically achieved by ITDs that far exceed this range (Fig. 1c3). More specifically, the preferred ITD of a cell depends on the best frequency (BF), irrespective of the head size of the species studied, and on average preferred ITDs increase with decreasing BF (Brand et al. 2002; Pecka et al. 2008; Middlebrooks et al. 1994; McAlpine et al. 2001; Hancock and Delgutte 2004; Werner-Reiss and Groh 2008). These data thus refuted the long-standing idea of the MSO as a distributed labeled-line encoder of azimuthal space in which preferred ITDs in each frequency band are distributed within the physiological range (or even clustered around the midline).

2.2 *Early Representation of Binaural Cues*

The fact that ITD functions of MSO neurons show very broad spatial tuning (essentially linear modulation between -90° and 90°) that is stereotypical within a given spectral band, stimulated the idea of hemispheric, oppositely coding channels on each side of the brain that might be compared at later stages of the pathway (McAlpine et al. 2001; Hancock and Delgutte 2004; Harper and McAlpine 2004; Stecker and Middlebrooks 2003; Pulkki and Hirvonen 2009) (Fig. 3a). This concept relies on the idea that similar activity levels in both channels should encode sound-source position at the midline, such that a relative increase in activity in one of the two brain hemispheres would indicate a corresponding contralateral location with respect to the more active brain hemisphere. This strategy might well optimize coding efficiency in animals with small head sizes, but might be sub-optimal for animals with larger heads and high-frequency hearing, owing to the increasing ambiguity of spike-rate-to-ITD mapping (Goodman et al. 2013; Harper et al. 2014). Such ambiguity could, however, be eliminated by making use of additional information from other ITD-sensitive channels (such as the low-frequency neurons in the LSO) during the decoding (Lüling et al. 2011; Day and Delgutte 2013; Goodman et al. 2013; Benichoux et al. 2015). In this framework it is important to keep in mind that the primary function of the MSO is the detection of ITDs. Consequently, the ITD-response tuning of MSO neurons to isolated sounds in experimental settings primarily captures this mechanistic function.

As we will see next, neuronal coding of space in the MSO and at downstream processing stages is strongly modulated by context, and these modulated representations are likely to reflect decoding aspects of perceptual sound localization.

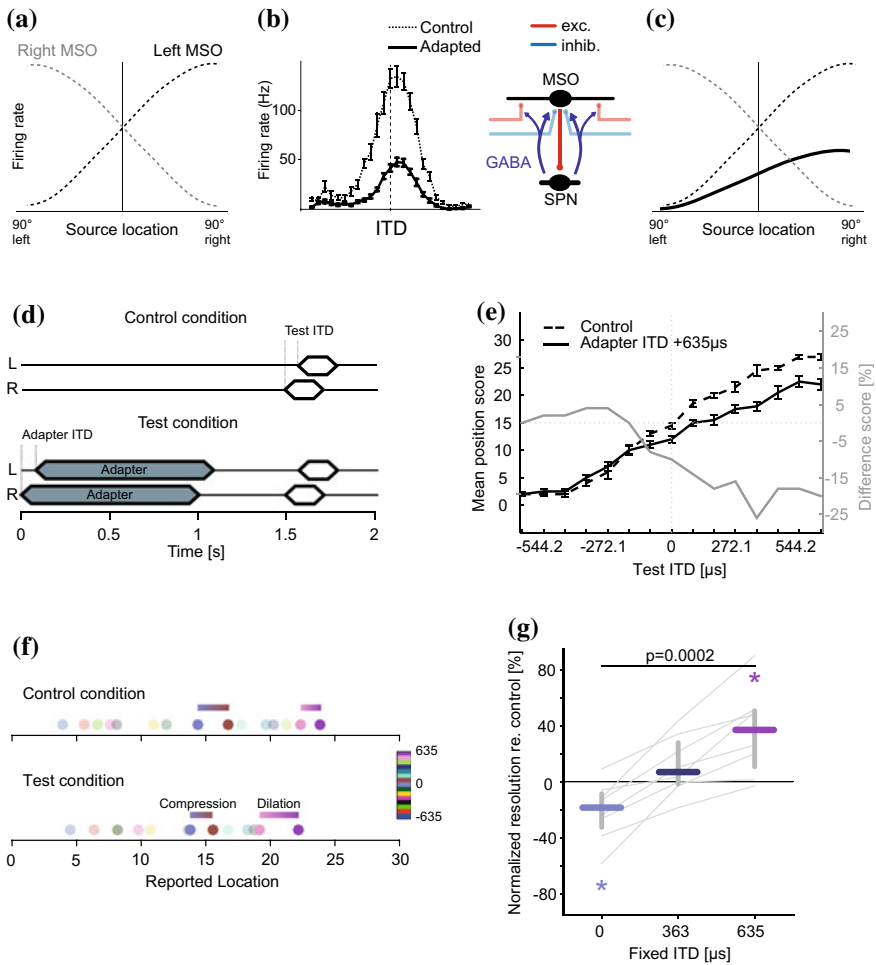
3 Context-Dependent Modulation of Neuronal and Perceptual-Space Representation

3.1 *Sub-Cortical Processing and Bottom-Up Modulation of Spatial Coding*

While the models for possible ITD representation are manifold, adaptation paradigms with human subjects not only strongly imply that some sort of hemispheric population code underlies sound localization in humans/primates, but have also highlighted the fact that prior stimulation influences subsequent spatial perception (Phillips and Hall 2005; Vigneault-MacLean et al. 2007). A number of studies have further demonstrated a dependency of the perceived location of a sound source on the spectral and spatial properties of the stimulus and their temporal profile (Kashino and Nishida 1998; Carlile et al. 2001; Getzmann 2004; Dahmen et al. 2010; Maier et al. 2012; Kopco et al. 2017). We will next review sub-cortical mechanisms and pathways for context-dependent sound localization that already correlated with perceptual phenomena in human listeners. We will then turn to more complex phenomena of context-dependent spatial hearing in real-world environments and discuss potential mechanisms.

It has become apparent that dynamic adaptation mechanisms act on spatial processing in the brainstem, including the MSO (Stange et al. 2013) and LSO (Magnusson et al. 2008). The response levels of individual neurons are negatively correlated with their prior spiking activity, which is typical of the concept of *gain modulation*, i.e. the dynamic adjustment of neuronal activity depending on its prior activity level. Notably, this modulation can last for seconds, as it is mediated by the inhibitory transmitter gamma-aminobutyric acid (GABA) and associated slow receptor-subtype signaling (GABAB). The consequences of these activity-dependent rate adaptations on binaural cue encoding have been studied in detail in the MSO (Fig. 3b): Here, GABAergic inhibition leads to an overall decrease in firing, the degree of which is indirectly proportional to the prior activity level of the cell. The ITD that elicits relative peak response rates of a particular neuron is not altered (i.e. there is no shift in the preferred ITD on the single cell level), but these modulations have significant consequences at the level of hemispheric population coding because of the activity-dependent nature of the adaptation. For example, a strongly lateralized sound source will generate unequal adaptation in the two hemispheres, with pronounced rate adaptation only in the contralateral channel. Accordingly, this hemispheric asymmetry should shift the perceived location of a subsequently presented sound source (Fig. 3c). As noted above, Phillips and colleagues first directly tested this hypothesis in a num-

ber of psychophysical paradigms, and were able to demonstrate a pronounced shift in the perceived location of sound sources after prior presentation of a lateralized adapter in human listeners (Phillips and Hall 2005; Vigneault-MacLean et al. 2007) (Fig. 3d,e) (Lingner et al. 2018). Grothe and colleagues have demonstrated (Stange et al. 2013) that GABAB-mediated rate adaptation in the MSO is already sufficient to explain these shifts in human perception. The primary function of these perceptual shifts seems to be the relative segregation of the adapting and the subsequent sound source, as the reported shifts in location are directed away from the adapter location, i.e. near the adapter location, the perceived distance between the sound sources is increased relative to the actual distance (Kashino and Nishida 1998; Vigneault-MacLean et al. 2007; Kopco et al. 2017) (Fig. 3f). This interpretation is supported by the finding that the presence of adapting sound sources increases spatial resolution



◀**Fig. 3** Context-dependent spatial-coding results in focal changes in separability. **a** Schematic of the hemispheric population coding principle. **b** Firing rates of MSO populations are modulated by recent stimulus history via a negative feedback loop, **c** resulting in hemispherically specific (i.e. asymmetric) adaptation. Conventions as in (b). **d** Stimulus paradigm to determine the intracranial perception of target tones without and with adapter. **e** Position scores (**0...most left** and **30...most right**) without and with adapter (**dashed and solid black line**) as a function of test ITD for one exemplary subject (mean \pm S.E.M.). The difference score (difference in position scores for each test ITD between control and test condition) is plotted in grey. Deviations from zero indicate a perceptual shift due to a preceding adapter (**negative...shift to the left, positive...shift to the right**). **f Upper panel:** Distribution of reported location values for the 15 test ITDs during the control condition for a representative subject. The resolvability of nearby test ITDs can be approximated by the difference in the corresponding location values—indicated by **horizontal bars**. **Lower panel:** Location values for the test condition for the same subject show pronounced shifts in the hemisphere ipsilateral to the adapter ITD, yet only marginal shifts around midline. These non-uniform shifts result in a compression, i.e. decreased differences between location values around midline (compare horizontal bar lengths near 0 μ s ITD). Simultaneously the perception of auditory space close to the adapter ITD is dilated, indicated by increased differences between location values for nearby test ITDs (compare horizontal bar lengths near 635 μ s ITD). **g** Median normalized resolution (re. control) across 8 subjects—**Colored bars:** **Grey lines** display individual subjects. Positive values indicate improved resolution, whereas negative values indicate deteriorated resolution due to the preceding adapter. The listeners' resolution increased significantly for locations closer to the adapter ($p = 0.00019$, Friedman test). Asterisks represent significantly altered resolution for ITD positions close to the adapter position and at 0 μ s, respectively ($p < 0.05$, two-sided Wilcoxon signed rank test). Reproduced from Lingner et al. (2018), copyright with the authors

at the adapter position, and likewise decreases resolution at positions further away from the adapter (Fig. 3g) (Getzmann 2004; Lingner et al. 2018).

Similar activity-dependent modulations have also been observed during ILD processing both psychophysically as well as in the LSO (Magnusson et al. 2008; Park et al. 2008) and midbrain (Dahmen et al. 2010). Together, these data strongly suggest that mammalian spatial coding serves to encode the relative separation of concurrent or subsequent sound sources already on the early levels of processing, which is in contradiction to the long-standing idea of maps of absolute auditory space in the brainstem.

While the present chapter emphasizes the implications of short-term plasticity on auditory space processing, there is also a large literature on long-term plasticity of spatial hearing investigating the effects of altered spectral and binaural cues as well as visual deprivation assays (Hofman et al. 1998; Zwiers et al. 2003; van Wanrooij and van Opstal 2005, 2007; Keating and King 2015). These experiments demonstrated a high capacity to relearn or form new associations between spatial cues and perceived locations based on recent experience (weeks to days) and cue reliability. These associations are likely to be formed at levels downstream to the detector neurons in the brainstem.

3.2 *Spatial Representations During Auditory Scene Analysis*

This emphasis on the relative separability of sound sources in mammalian neural processing already on the interaural detector level could facilitate spatial coding in complex listening environments with multiple sound sources. Particularly, dynamic spatial coding appears to be advantageous for auditory scene analysis (ASA), that is, the capacity of the brain to deconvolve the complex mixtures of sound waves stemming from multiple sound sources and to group them into distinct streams of information according to the origins. Traditionally, the primary Auditory Cortex (A1), the downstream cortical areas, and their pathways, were the dominant sites of research on neuronal correlates of cognitive effects in ASA (Micheyl et al. 2007; Tsunada and Cohen 2014; Christison-Lagay et al. 2015). Yet, it has become clear in recent years that ASA is not the product of the processing of only a small local population of neurons in a specific region of the brain. Rather, many nuclei along the ascending auditory pathway are involved in the gradual transformation of *low-level* (i.e. single-feature based) to perceptual representations of auditory events, see, Shamma and Micheyl (2010); Nelken et al. (2014); Clarke and Geiser (2015); Osmanski and Wang (2015) —Fig. 4. There is accumulating evidence for the significance of early processing even upstream of A1 for the generation of many fundamental functions of ASA such as feature grouping and contextual feature streaming (Snyder and Alain 2007; Pressnitzer et al. 2008; Shamma and Micheyl 2010).

For example, the afore-mentioned stimulus-history-dependent dynamic-range adaptations in spatial tuning in the brainstem (Magnusson et al. 2008; Stange et al. 2013) and midbrain (Dahmen et al. 2010) can be regarded as initial mechanisms for selective stream formation. Such early contribution might be the product of the unique hierarchical processing structure along the auditory pathway, which—as has been mentioned in the Introduction—is inherently different as compared to other sensory systems such as the visual one. Since the cochlea decomposes all incoming sounds into distinct frequency channels, so-called *critical bands*, downstream neuronal feature analysis is performed independently for each of these bands. It follows that any spatial correlations that could be directly exploited, are missing. Rather, the perceptual grouping of sounds which (presumably) belong to the same *auditory object* must be neurally reconstructed based on common features across frequency channels—so-called *grouping cues*. Next to a common harmonic structure of sounds, a crucial grouping cue is the common position in space. Interestingly, activity associated with stream segregation in the auditory cortex seems to be similar for spatial and pitch related cues (Schadwinkel and Gutschalk 2010, 2011), and the role of spatial cues for sound-identity encoding is crucial in the presence of multiple sound sources—compare Maddox et al. (2012).

3.3 Cortical Spatial Representations under Complex Conditions

As explained above, spatial-cue analysis of an auditory stimulus is performed already in the brainstem. It is then followed up in the IC, which is the “midbrain hub” where information from various cues converges (Grothe et al. 2010) (Fig. 4). Spatial tuning based on ITDs has been shown to be further transformed between IC and A1, potentially to facilitate listening in complex scenarios (Belliveau et al. 2014; Yao et al. 2015). Yet the role of cortical processing in spatial hearing in complex settings and during ASA is highly multifaceted (Bizley and Cohen 2013; Lewicki et al. 2014). The following discussion will therefore be limited selectively to findings of early facultative representations and focus on how spatial information is neurally represented in A1 under dynamic conditions or in complex acoustic environments (e.g., multiple sources). Most electrophysiological studies (Ahissar et al. 1992; Stecker et al. 2005; Woods et al. 2006; Werner-Reiss and Groh 2008; Yao et al. 2015), including such from humans (Salminen et al. 2010, 2018), indicated that neurons in A1 are broadly tuned to sound-source location, where the spike rate is linearly modulated across

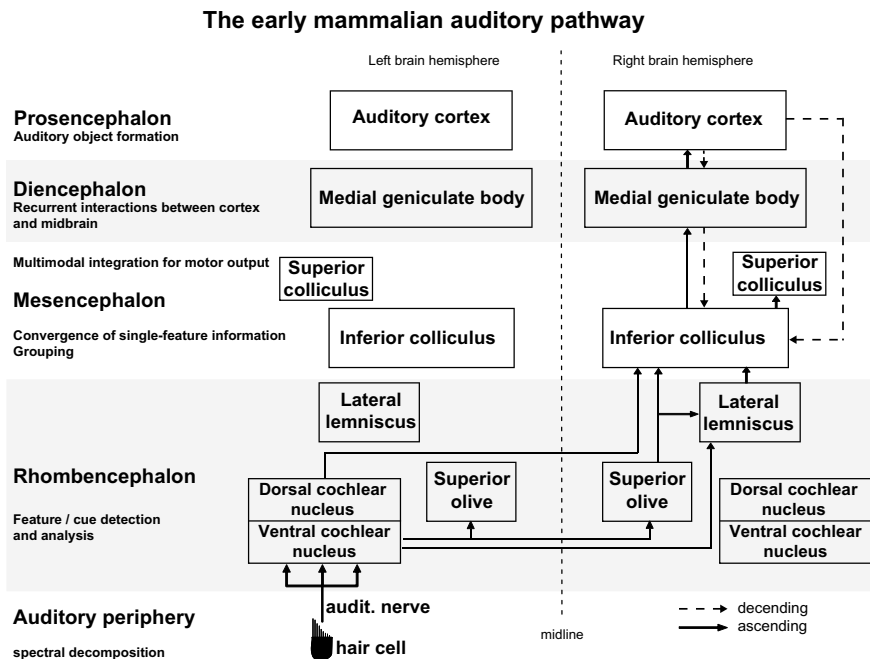


Fig. 4 Construction of auditory objects along the sub-cortical mammalian auditory pathway. Simplified circuit diagram of the ascending auditory pathway, including descending projections from the cortex), and the potential relationships of individual stages of processing with auditory scene analysis

the entire frontal field with maximal rates for contralateral positions. This representation is remarkably consistent with what can be found already at the level of the IC and even the brainstem (differing only in additional information being present in response timing in A1, (Stecker and Middlebrooks 2003)). This suggests a high level of redundancy in the evaluation processes. However, the vast majority of relevant studies was performed in anesthetized preparations and/or used rather simple stimulation paradigms that lacked the complexity of natural environments.

Consequently, applying a multi-source environment as a simplistic approximation to a complex *auditory scene* already seems to invoke mechanisms that lead to rather specific spatial representations, reminiscent of aspects of ASA. Multiple recent studies point towards a progressive refinement of auditory information processing to facilitate hearing in complex settings, that is, in scenes that involve multiple sound sources or a noise background. Even in the anesthetized brain, signal-to-noise ratio improves on the way from the auditory nerve via midbrain to cortical representations (Rabinowitz et al. 2013; Willmore et al. 2016). Similar improvement in the separability of fore- and background, as well as level-tolerance, has been found with regard to spatial information between the IC and the A1—Yao et al. (2015).

An interesting hypothesis about spatial coding has recently been put forward by Town et al. (2017), suggesting that a minority of neurons in the A1 might represent sound locations in an allocentric manner, that is, independent of the listeners own position or head orientation. This is indeed what is indispensable for the orientation of listeners in complex scenarios.

It is a defining motive of A1 that its spatial sensitivity can be modulated by behavioral task requirements (Benson et al. 1981; Lee and Middlebrooks 2011; Salminen et al. 2012). This could be interpreted as the reflection of transient and context-specific processing motives in the ASA process—rather than universal representations of feature selectivity.

Lee and Middlebrooks (2011) demonstrated that neuronal spatial tuning in the A1 of cats emerged during active localization but vanished in passive hearing settings or during a spectral detection task. The behavioral relevance of spatial information is often conveyed by visual signal, and it is known that many neurons in A1 are sensitive to combinations of auditory and visual stimuli (Kayser et al. 2008; Yau et al. 2015). Moreover, it has been shown that combined stimulation of both sensory modalities can lead to an enhancement of auditory information, particularly about location, both in the IC (Bizley and King 2008) and in the A1 (Kayser et al. 2010). Strikingly, in A1 this enhancement is mediated by a reduction in spiking, hence resulting in a more sparse and more efficient coding (Ghazanfar and Lemus 2010).

Multi-modal stimulation is generally crucial for goal- or relevance-driven analysis of features, as it can provide valuable information for priming (Noppeney et al. 2008; Diehl and Romanski 2014; Altieri et al. 2015) and inference (Beck et al. 2012). For example, visual cues improve the conclusion about the most likely location of a sound source. Accordingly, recent studies showed strong effects of visually induced relevance regarding the response magnitude in both the A1 of awake gerbils (Kobayasi et al. 2013) and the human auditory cortex (van Wassenhove and Grzeczkowski 2015).

Interestingly, compared to other modalities, efferent projections from the cortex to the midbrain are very pronounced—see Bajo and Moore (2005); Suga (2008); Stebbings et al. (2014) and Fig. 4. These are involved in learning-induced auditory plasticity of spatial cues as well as in sound-induced innate behavior (Bajo et al. 2010; Xiong et al. 2015). Inactivating this cortical feedback alters the patterns of representations of concurrent stimuli in IC, which are thought to be involved in streaming (Nakamoto et al. 2010, 2015). For these reasons, the recurring interactions between IC and A1 are also associated with mechanisms of attention and/or expectation-based streaming of sensory information (Middlebrooks and Bremen 2013; Malmierca et al. 2015). This hypothesis is further supported by human functional-imaging data, showing time-locked activity in the IC and the A1, synchronized with epochs of auditory-stream formation (Schadwinkel and Gutschalk 2011).

4 Conclusion

In summary, spatial processing in the mammalian (including the human) auditory system is characterized by a lack of topography as can be found in other sensory systems and, further, by a high degree of contextual modulation. Together, these motives suggest that not an absolute representation, but a focused separability of the most relevant sources (i.e. perceptual objects) in complex environments is the primary objective of auditory space perception.

Acknowledgements The authors are indebted to two anonymous reviewers for constructive comments and suggestions.

References

- Ahissar, M., E. Ahissar, H. Bergman, and E. Vaadia. 1992. Encoding of sound-source location and movement: activity of single neurons and interactions between adjacent neurons in the monkey auditory cortex. *Journal of Neurophysiology* 67 (1): 203–215. <https://doi.org/10.1152/jn.1992.67.1.203>.
- Allin, E.F. 1975. Evolution of the mammalian middle ear. *Journal of morphology* 147 (4): 403–437. <https://doi.org/10.1002/jmor.1051470404>.
- Altieri, N., R.A. Stevenson, M.T. Wallace, and M.J. Wenger. 2015. Learning to associate auditory and visual stimuli: behavioral and neural mechanisms. *Brain Topography* 28 (3): 479–493. <https://doi.org/10.1007/s10548-013-0333-7>.
- Ashida, G., J. Kretzberg, and D.J. Tollin. 2016. Roles for Coincidence Detection in Coding Amplitude-Modulated Sounds. *PLoS computational biology* 12 (6): e1004997. <https://doi.org/10.1371/journal.pcbi.1004997>.
- Bajo, V.M., and D.R. Moore. 2005. Descending projections from the auditory cortex to the inferior colliculus in the gerbil, *Meriones unguiculatus*. *The Journal of Comparative Neurology* 486 (2): 101–116. <https://doi.org/10.1002/cne.20542>.

- Bajo, V.M., F.R. Nodal, D.R. Moore, and A.J. King. 2010. The descending corticocollicular pathway mediates learning-induced auditory plasticity. *Nature Neuroscience* 13 (2): 253–260. <https://doi.org/10.1038/nn.2466>.
- Beck, J.M., W.J. Ma, X. Pitkow, P.E. Latham, and A. Pouget. 2012. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* 74 (1): 30–39. <https://doi.org/10.1016/j.neuron.2012.03.016>.
- Belliveau, L.A.C., D.R. Lyamzin, and N.A. Lesica. 2014. The neural representation of interaural time differences in gerbils is transformed from midbrain to cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 34 (50): 16796–16808. <https://doi.org/10.1523/JNEUROSCI.2432-14.2014>.
- Benichoux, V., B. Fontaine, T.P. Franken, S. Karino, P.X. Joris, and R. Brette. 2015. Neural tuning matches frequency-dependent time differences between the ears. *eLife* 4. <https://doi.org/10.7554/eLife.06072>.
- Benson, D.A., R.D. Hienz, and M.H. Goldstein. 1981. Single-unit activity in the auditory cortex of monkeys actively localizing sound sources: spatial tuning and behavioral dependency. *Brain Research* 219 (2): 249–267.
- Bizley, J.K., and Y.E. Cohen. 2013. The what, where and how of auditory-object perception. *Nature Reviews Neuroscience* 14 (10): 693–707. <https://doi.org/10.1038/nrn3565>.
- Bizley, J.K., and A.J. King. 2008. Visual-auditory spatial processing in auditory cortical neurons. *Brain Research* 1242: 24–36. <https://doi.org/10.1016/j.brainres.2008.02.087>.
- Boudreau, J.C., and C. Tsuchitani. 1968. Binaural interaction in the cat superior olive S segment. *Journal of neurophysiology* 31 (3): 442–454.
- Brand, A., O. Behrend, T. Marquardt, D. McAlpine, and B. Grothe. 2002. Precise inhibition is essential for microsecond interaural time difference coding. *Nature* 417 (6888): 543–547. <https://doi.org/10.1038/417543a>.
- Carlile, S., S. Hyams, and S. Delaney. 2001. Systematic distortions of auditory space perception following prolonged exposure to broadband noise. *The Journal of the Acoustical Society of America* 110 (1): 416–424.
- Christison-Lagay, K.L., A.M. Gifford, and Y.E. Cohen. 2015. Neural correlates of auditory scene analysis and perception. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology* 95 (2): 238–245. <https://doi.org/10.1016/j.ijpsycho.2014.03.004>.
- Clack, J.A. 1997. The evolution of tetrapod ears and the fossil record. *Brain, behavior and evolution* 50 (4): 198–212.
- Clarke, S., and E. Geiser. 2015. Roaring lions and chirruping lemurs: How the brain encodes sound objects in space. *Neuropsychologia* 75: 304–313. <https://doi.org/10.1016/j.neuropsychologia.2015.06.012>.
- Couchman, K., B. Grothe, and F. Felmy. 2010. Medial superior olivary neurons receive surprisingly few excitatory and inhibitory inputs with balanced strength and short-term dynamics. *The Journal of neuroscience: the official journal of the Society for Neuroscience* 30 (50): 17111–17121. <https://doi.org/10.1523/JNEUROSCI.1760-10.2010>.
- Dahmen, J.C., P. Keating, F.R. Nodal, A.L. Schulz, and A.J. King. 2010. Adaptation to stimulus statistics in the perception and neural representation of auditory space. *Neuron* 66 (6): 937–948. <https://doi.org/10.1016/j.neuron.2010.05.018>.
- Davis, K.A., R. Ramachandran, and B.J. May. 2003. Auditory processing of spectral cues for sound localization in the inferior colliculus. *Journal of the Association for Research in Otolaryngology: JARO* 4 (2): 148–163. <https://doi.org/10.1007/s10162-002-2002-5>.
- Day, M.L., and B. Delgutte. 2013. Decoding sound source location and separation using neural population activity patterns. *The Journal of neuroscience: the official journal of the Society for Neuroscience* 33 (40): 15837–15847. <https://doi.org/10.1523/JNEUROSCI.2034-13.2013>.
- Diehl, M.M., and L.M. Romanski. 2014. Responses of prefrontal multisensory neurons to mismatching faces and vocalizations. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 34 (34): 11233–11243. <https://doi.org/10.1523/JNEUROSCI.5168-13.2014>.

- Erulkar, S.D. 1972. Comparative aspects of spatial localization of sound. *Physiological reviews* 52 (1): 237–360.
- Finlayson, P.G., and D.M. Caspary. 1991. Low-frequency neurons in the lateral superior olive exhibit phase-sensitive binaural inhibition. *Journal of neurophysiology* 65 (3): 598–605.
- Fischer, L., C. Leibold, and F. Felmy. 2018. Resonance properties in auditory brainstem neurons. *Frontiers in Cellular Neuroscience* 12: 8.
- Franken, T.P., P. Bremen, and P.X. Joris. 2014. Coincidence detection in the medial superior olive: mechanistic implications of an analysis of input spiking patterns. *Frontiers in Neural Circuits* 8: 42. <https://doi.org/10.3389/fncir.2014.00042>.
- Galambos, R., J. Schwartzkopff, and A. Rupert. 1959. Microelectrode study of superior olivary nuclei. *The American Journal of Physiology* 197: 527–536.
- Getzmann, S. 2004. Spatial discrimination of sound sources in the horizontal plane following an adapter sound. *Hearing Research* 191 (1–2): 14–20. <https://doi.org/10.1016/j.heares.2003.12.020>.
- Ghazanfar, A.A., and L. Lemus. 2010. Multisensory integration: vision boosts information through suppression in auditory cortex. *Current Biology: CB* 20 (1): R22–23. <https://doi.org/10.1016/j.cub.2009.11.046>.
- Goodman, D.F., V. Benichoux, and R. Brette. 2013. Decoding neural responses to temporal cues for sound localization. *eLife* 2: e01312. <https://doi.org/10.7554/eLife.01312>.
- Grothe, B. 2000. The evolution of temporal processing in the medial superior olive, an auditory brainstem structure. *Progress in Neurobiology* 61 (6): 581–610.
- Grothe, B., and M. Pecka. 2014. The natural history of sound localization in mammals—a story of neuronal inhibition. *Frontiers in Neural Circuits* 8: 116. <https://doi.org/10.3389/fncir.2014.00116>.
- Grothe, B., M. Pecka, and D. McAlpine. 2010. Mechanisms of sound localization in mammals. *Physiological Reviews* 90 (3): 983–1012. <https://doi.org/10.1152/physrev.00026.2009>.
- Hancock, K.E., and B. Delgutte. 2004. A physiologically based model of interaural time difference discrimination. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 24(32): 7110–7117. <http://www.ncbi.nlm.nih.gov/pubmed/15306644>, <https://doi.org/10.1523/JNEUROSCI.0762-04.2004>.
- Harnischfeger, G., G. Neuweiler, and P. Schlegel. 1985. Interaural time and intensity coding in superior olivary complex and inferior colliculus of the echolocating bat *Molossus ater*. *Journal of Neurophysiology* 53 (1): 89–109.
- Harper, N.S., and D. McAlpine. 2004. Optimal neural population coding of an auditory spatial cue. *Nature* 430 (7000): 682–686. <https://doi.org/10.1038/nature02768>.
- Harper, N.S., B.H. Scott, M.N. Semple, and D. McAlpine. 2014. The neural code for auditory space depends on sound frequency and head size in an optimal manner. *PloS One* 9 (11): e108154. <https://doi.org/10.1371/journal.pone.0108154>.
- Hofman, P.M., J.G. Van Riswick, and A.J. Van Opstal. 1998. Relearning sound localization with new ears. *Nature Neuroscience* 1 (5): 417–421.
- Hylander, W.L., and A.W. Crompton. 1986. Jaw movements and patterns of mandibular bone strain during mastication in the monkey *Macaca fascicularis*. *Archives of Oral Biology* 31 (12): 841–848.
- Imig, T.J., N.G. Bibikov, P. Poirier, and F.K. Samson. 2000. Directionality derived from pinna-cue spectral notches in cat dorsal cochlear nucleus. *Journal of Neurophysiology* 83 (2): 907–925. <https://doi.org/10.1152/jn.2000.83.2.907>.
- Jeffress, L.A. 1948. A place theory of sound localization. *Journal of Comparative and Physiological Psychology* 41 (1): 35–39.
- Joris, P., and T.C.T. Yin. 2007. A matter of time: internal delays in binaural processing. *Trends in Neurosciences* 30 (2): 70–78. <https://doi.org/10.1016/j.tins.2006.12.004>.
- Kashino, M., and S. Nishida. 1998. Adaptation in the processing of interaural time differences revealed by the auditory localization aftereffect. *The Journal of the Acoustical Society of America* 103 (6): 3597–3604.

- Kaysers, C., N.K. Logothetis, and S. Panzeri. 2010. Millisecond encoding precision of auditory cortex neurons. *Proceedings of the National Academy of Sciences of the United States of America* 107 (39): 16976–16981. <https://doi.org/10.1073/pnas.1012656107>.
- Kaysers, C., C.I. Petkov, and N.K. Logothetis. 2008. Visual modulation of neurons in auditory cortex. *Cerebral Cortex* (New York, N.Y.: 1991) 18(7): 1560–1574. <https://doi.org/10.1093/cercor/bhm187>.
- Keating, P., and A.J. King. 2013. Developmental plasticity of spatial hearing following asymmetric hearing loss: context-dependent cue integration and its clinical implications. *Frontiers in Systems Neuroscience* 7: 123. <https://doi.org/10.3389/fnsys.2013.00123>.
- Keating, P., and A.J. King. 2015. Sound localization in a changing world. *Current Opinion in Neurobiology* 35: 35–43. <https://doi.org/10.1016/j.conb.2015.06.005>.
- Keating, P., F.R. Nodal, K. Gananandan, A.L. Schulz, and A.J. King. 2013. Behavioral sensitivity to broadband binaural localization cues in the ferret. *Journal of the Association for Research in Otolaryngology: JARO* 14 (4): 561–572. <https://doi.org/10.1007/s10162-013-0390-3>.
- Keating, P., O. Rosenior-Patten, J.C. Dahmen, O. Bell, and A.J. King. 2016. Behavioral training promotes multiple adaptive processes following acute hearing loss. *eLife* 5: e12264. <https://doi.org/10.7554/eLife.12264>.
- Kobayashi, K.I., Y. Suwa, and H. Riquimaroux. 2013. Audiovisual integration in the primary auditory cortex of an awake rodent. *Neuroscience Letters* 534: 24–29. <https://doi.org/10.1016/j.neulet.2012.10.056>.
- Kopco, N., G. Andrejkova, V. Best, and B. Shinn-Cunningham. 2017. Streaming and sound localization with a preceding distractor. *The Journal of the Acoustical Society of America* 141(4): EL331. <https://doi.org/10.1121/1.4979167>.
- Lee, C.-C., and J.C. Middlebrooks. 2011. Auditory cortex spatial sensitivity sharpens during task performance. *Nature Neuroscience* 14 (1): 108–114. <https://doi.org/10.1038/nn.2713>.
- Lewicki, M.S., B.A. Olshausen, A. Surllykke, and C.F. Moss. 2014. Scene analysis in the natural environment. *Frontiers in Psychology* 5: 199. <https://doi.org/10.3389/fpsyg.2014.00199>.
- Lingner, A., M. Pecka, C. Leibold., and B. Grothe. 2018. A novel concept for dynamic adjustment of auditory space. *Scientific Reports* 8: 8335.
- Lüling, H., I. Siveke, B. Grothe, and C. Leibold. 2011. Frequency-invariant representation of interaural time differences in mammals. *PLoS Computational Biology* 7(3): e1002013. <https://doi.org/10.1371/journal.pcbi.1002013>.
- Maddox, R.K., C.P. Billimoria, B.P. Perrone, B.G. Shinn-Cunningham, and K. Sen. 2012. Competing sound sources reveal spatial effects in cortical processing. *PLoS Biology* 10 (5): e1001319. <https://doi.org/10.1371/journal.pbio.1001319>.
- Magnusson, A.K., T.J. Park, M. Pecka, B. Grothe, and U. Koch. 2008. Retrograde GABA signaling adjusts sound localization by balancing excitation and inhibition in the brainstem. *Neuron* 59 (1): 125–137. <https://doi.org/10.1016/j.neuron.2008.05.011>.
- Maier, J.K., P. Hehrmann, N.S. Harper, G.M. Klump, D. Pressnitzer, and D. McAlpine. 2012. Adaptive coding is constrained to midline locations in a spatial listening task. *Journal of Neurophysiology* 108 (7): 1856–1868. <https://doi.org/10.1152/jn.00652.2011>.
- Malmierca, M.S., L.A. Anderson, and F.M. Antunes. 2015. The cortical modulation of stimulus-specific adaptation in the auditory midbrain and thalamus: a potential neuronal correlate for predictive coding. *Frontiers in Systems Neuroscience* 9: 19. <https://doi.org/10.3389/fnsys.2015.00019>.
- McAlpine, D., D. Jiang, and A.R. Palmer. 2001. A neural code for low-frequency sound localization in mammals. *Nature Neuroscience* 4 (4): 396–401. <https://doi.org/10.1038/86049>.
- Mendonça, C. 2014. A review on auditory space adaptations to altered head-related cues. *Frontiers in Neuroscience* 8: 219. <https://doi.org/10.3389/fnins.2014.00219>.
- Micheyl, C., R.P. Carlyon, A. Gutschalk, J.R. Melcher, A.J. Oxenham, J.P. Rauschecker, B. Tian, and E. Courtenay Wilson. 2007. The role of auditory cortex in the formation of auditory streams. *Hearing Research* 229 (1–2): 116–131. <https://doi.org/10.1016/j.heares.2007.01.007>.

- Middlebrooks, J.C., and P. Bremen. 2013. Spatial stream segregation by auditory cortical neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 33 (27): 10986–11001. <https://doi.org/10.1523/JNEUROSCI.1065-13.2013>.
- Middlebrooks, J.C., A.E. Clock, L. Xu, and D.M. Green. 1994. A panoramic code for sound location by cortical neurons. *Science* (New York, N.Y.) 264(5160): 842–844.
- Myoga, M.H., S. Lehnert, C. Leibold, F. Felmy, and B. Grothe. 2014. Glycinergic inhibition tunes coincidence detection in the auditory brainstem. *Nature Communications* 5: 3790. <https://doi.org/10.1038/ncomms4790>.
- Nakamoto, K.T., T.M. Shackleton, D.A. Magezi, and A.R. Palmer. 2015. A function for binaural integration in auditory grouping and segregation in the inferior colliculus. *Journal of Neurophysiology* 113 (6): 1819–1830. <https://doi.org/10.1152/jn.00472.2014>.
- Nakamoto, K.T., T.M. Shackleton, and A.R. Palmer. 2010. Responses in the inferior colliculus of the guinea pig to concurrent harmonic series and the effect of inactivation of descending controls. *Journal of Neurophysiology* 103 (4): 2050–2061. <https://doi.org/10.1152/jn.00451.2009>.
- Nelken, I., J. Bizley, S.A. Shamma, and X. Wang. 2014. Auditory cortical processing in real-world listening: the auditory system going real. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 34 (46): 15135–15138. <https://doi.org/10.1523/JNEUROSCI.2989-14.2014>.
- Noppeney, U., O. Josephs, J. Hocking, C.J. Price, and Friston, K.J. 2008. The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex* (New York, N.Y.: 1991) 18(3): 598–609. <https://doi.org/10.1093/cercor/bhm091>.
- Oertel, D., and E.D. Young. 2004. What’s a cerebellar circuit doing in the auditory system? *Trends in Neurosciences* 27 (2): 104–110. <https://doi.org/10.1016/j.tins.2003.12.001>.
- Osmanski, M.S., and X. Wang. 2015. Behavioral dependence of auditory cortical responses. *Brain Topography* 28 (3): 365–378. <https://doi.org/10.1007/s10548-015-0428-4>.
- Park, T.J., A. Brand, U. Koch, M. Ikebuchi, and B. Grothe. 2008. Dynamic changes in level influence spatial coding in the lateral superior olive. *Hearing Research* 238 (1–2): 58–67. <https://doi.org/10.1016/j.heares.2007.10.009>.
- Park, T.J., B. Grothe, G.D. Pollak, G. Schuller, and U. Koch. 1996. Neural delays shape selectivity to interaural intensity differences in the lateral superior olive. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 16 (20): 6554–6566.
- Pecka, M., A. Brand, O. Behrend, and B. Grothe. 2008. Interaural time difference processing in the mammalian medial superior olive: the role of glycinergic inhibition. *Journal of Neuroscience* 28 (27): 6914–6925. <https://doi.org/10.1523/JNEUROSCI.1660-08.2008>.
- Phillips, D.P., and S.E. Hall. 2005. Psychophysical evidence for adaptation of central auditory processors for interaural differences in time and level. *Hearing Research* 202 (1–2): 188–199. <https://doi.org/10.1016/j.heares.2004.11.001>.
- Pressnitzer, D., M. Sayles, C. Micheyl, and I.M. Winter. 2008. Perceptual organization of sound begins in the auditory periphery. *Current Biology: CB* 18 (15): 1124–1128. <https://doi.org/10.1016/j.cub.2008.06.053>.
- Pulkki, V., and T. Hirvonen. 2009. Functional count-comparison model for binaural decoding. *Acta Acustica United with Acustica* 95 (5): 883–900.
- Rabinowitz, N.C., B.D.B. Willmore, A.J. King, and J.W.H. Schnupp. 2013. Constructing noise-invariant representations of sound in the auditory pathway. *PLoS Biology* 11 (11): e1001710. <https://doi.org/10.1371/journal.pbio.1001710>.
- Rosowski, J.J. 1991. The effects of external- and middle-ear filtering on auditory threshold and noise-induced hearing loss. *The Journal of the Acoustical Society of America* 90 (1): 124–135.
- Salminen, N.H., S.J. Jones, G.B. Christianson, T. Marquardt, and D. McAlpine. 2018. A common periodic representation of interaural time differences in mammalian cortex. *NeuroImage* 167: 95–103. <https://doi.org/10.1016/j.neuroimage.2017.11.012>.
- Salminen, N.H., H. Tiitinen, and P.J.C. May. 2012. Auditory spatial processing in the human cortex. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry* 18 (6): 602–612. <https://doi.org/10.1177/1073858411434209>.

- Salminen, N.H., H. Tiitinen, S. Yrttiaho, and P.J.C. May. 2010. The neural code for interaural time difference in human auditory cortex. *The Journal of the Acoustical Society of America* 127(2): EL60–65. <https://doi.org/10.1121/1.3290744>.
- Schadwinkel, S., and A. Gutschalk. 2010. Activity associated with stream segregation in human auditory cortex is similar for spatial and pitch cues. *Cerebral Cortex* (New York, N.Y.: 1991) 20(12): 2863–2873. <https://doi.org/10.1093/cercor/bhq037>.
- Schadwinkel, S., and A. Gutschalk. 2011. Transient bold activity locked to perceptual reversals of auditory streaming in human auditory cortex and inferior colliculus. *Journal of Neurophysiology* 105 (5): 1977–1983. <https://doi.org/10.1152/jn.00461.2010>.
- Scott, L.L., P.J. Mathews, and N.L. Golding. 2005. Posthearing developmental refinement of temporal processing in principal neurons of the medial superior olive. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 25 (35): 7887–7895. <https://doi.org/10.1523/JNEUROSCI.1016-05.2005>.
- Shamma, S.A., and C. Micheyl. 2010. Behind the scenes of auditory perception. *Current Opinion in Neurobiology* 20 (3): 361–366. <https://doi.org/10.1016/j.conb.2010.03.009>.
- Snyder, J.S., and C. Alain. 2007. Toward a neurophysiological theory of auditory stream segregation. *Psychological Bulletin* 133 (5): 780–799. <https://doi.org/10.1037/0033-2909.133.5.780>.
- Stange, A., M.H. Myoga, A. Lingner, M.C. Ford, O. Alexandrova, F. Felmy, M. Pecka, I. Siveke, and B. Grothe. 2013. Adaptation in sound localization: from GABAB receptor-mediated synaptic modulation to perception. *Nature Neuroscience* 16 (12): 1840–1847. <https://doi.org/10.1038/nn.3548>.
- Stebbing, K.A., A.M.H. Lesicko, and D.A. Llano. 2014. The auditory corticocollicular system: molecular and circuit-level considerations. *Hearing Research* 314: 51–59. <https://doi.org/10.1016/j.heares.2014.05.004>.
- Stecker, G.C., I.A. Harrington, and J.C. Middlebrooks. 2005. Location coding by opponent neural populations in the auditory cortex. *PLoS Biology* 3 (3): e78. <https://doi.org/10.1371/journal.pbio.0030078>.
- Stecker, G.C., and J.C. Middlebrooks. 2003. Distributed coding of sound locations in the auditory cortex. *Biological Cybernetics* 89 (5): 341–349. <https://doi.org/10.1007/s00422-003-0439-1>.
- Suga, N. 2008. Role of corticofugal feedback in hearing. *Journal of Comparative Physiology A, Neuroethology, Sensory, Neural, and Behavioral Physiology* 194 (2): 169–183. <https://doi.org/10.1007/s00359-007-0274-2>.
- Tollin, D.J. 2003. The lateral superior olive: a functional role in sound source localization. *The Neuroscientist: a Review Journal Bringing Neurobiology, Neurology and Psychiatry* 9 (2): 127–143.
- Tollin, D.J., and T.C.T. Yin. 2005. Interaural phase and level difference sensitivity in low-frequency neurons in the lateral superior olive. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 25 (46): 10648–10657. <https://doi.org/10.1523/JNEUROSCI.1609-05.2005>.
- Town, S.M., W.O. Brimijoin, and J.K. Bizley. 2017. Egocentric and allocentric representations in auditory cortex. *PLoS Biology* 15 (6): e2001878. <https://doi.org/10.1371/journal.pbio.2001878>.
- Tsuchitani, C., and J.C. Boudreau. 1969. Stimulus level of dichotically presented tones and cat superior olive S-segment cell discharge. *The Journal of the Acoustical Society of America* 46 (4): 979–988.
- Tsunada, J., and Y.E. Cohen. 2014. Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits. *Frontiers in Neuroscience* 8: 161. <https://doi.org/10.3389/fnins.2014.00161>.
- van der Heijden, M., J.A.M. Lorteije, A. Plauka, M.T. Roberts, N.L. Golding, and J.G.G. Borst. 2013. Directional hearing by linear summation of binaural inputs at the medial superior olive. *Neuron* 78 (5): 936–948. <https://doi.org/10.1016/j.neuron.2013.04.028>.
- van Wanrooij, M.M., and A.J. van Opstal. 2005. Relearning sound localization with a new ear. *Journal of Neuroscience* 25 (22): 5413–5424.

- van Wanrooij, M.M., and A.J. van Opstal. 2007. Sound localization under perturbed binaural hearing. *Journal of Neurophysiology* 97 (1): 715–726.
- van Wassenhove, V., and L. Grzeczowski. 2015. Visual-induced expectations modulate auditory cortical responses. *Frontiers in Neuroscience* 9: 11. <https://doi.org/10.3389/fnins.2015.00011>.
- Vigneault-MacLean, B.K., S.E. Hall, and D.P. Phillips. 2007. The effects of lateralized adaptors on lateral position judgements of tones within and across frequency channels. *Hearing Research* 224 (1–2): 93–100. <https://doi.org/10.1016/j.heares.2006.12.001>.
- Werner-Reiss, U., and J.M. Groh. 2008. A rate code for sound azimuth in monkey auditory cortex: implications for human neuroimaging studies. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 28 (14): 3747–3758. <https://doi.org/10.1523/JNEUROSCI.5044-07.2008>.
- Willmore, B.D.B., O. Schoppe, A.J. King, J.W.H. Schnupp, and N.S. Harper. 2016. Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 36 (2): 280–289. <https://doi.org/10.1523/JNEUROSCI.2441-15.2016>.
- Woods, T.M., S.E. Lopez, J.H. Long, J.E. Rahman, and G.H. Recanzone. 2006. Effects of stimulus azimuth and intensity on the single-neuron activity in the auditory cortex of the alert macaque monkey. *Journal of Neurophysiology* 96 (6): 3323–3337. <https://doi.org/10.1152/jn.00392.2006>.
- Xiong, X.R., F. Liang, B. Zingg, X.-Y. Ji, L.A. Ibrahim, H.W. Tao, and L.I. Zhang. 2015. Auditory cortex controls sound-driven innate defense behaviour through corticofugal projections to inferior colliculus. *Nature Communications* 6: 7224. <https://doi.org/10.1038/ncomms8224>.
- Yao, J.D., P. Bremen, and J.C. Middlebrooks. 2015. Transformation of spatial sensitivity along the ascending auditory pathway. *Journal of Neurophysiology* 113 (9): 3098–3111. <https://doi.org/10.1152/jn.01029.2014>.
- Yau, J.M., G.C. DeAngelis, and D.E. Angelaki. 2015. “Dissecting neural circuits for multisensory integration and crossmodal processing” Philosophical Transactions of the Royal Society of London. *Series B, Biological Sciences* 370 (1677): 20140203. <https://doi.org/10.1098/rstb.2014.0203>.
- Young, E.D., G.A. Spirou, J.J. Rice, and H.F. Voigt. 1992. “Neural organization and responses to complex stimuli in the dorsal cochlear nucleus” Philosophical Transactions of the Royal Society of London. *Series B, Biological Sciences* 336 (1278): 407–413. <https://doi.org/10.1098/rstb.1992.0076>.
- Zwiers, M.P., A.J. Van Opstal, and G.D. Paige. 2003. Plasticity in human sound localization induced by compressed spatial vision. *Nature Neuroscience* 6 (2): 175–181.

Auditory Spatial Impression in Concert Halls



Tapio Lokki and Jukka Pätynen

Abstract This chapter discusses the acoustics of concert halls from the viewpoint of binaural perception. It explains how early reflections have a crucial role in the quality of sound, perceived dynamics, and timbre. In particular, the directions from which these reflections reach the listener are important for human spatial hearing. The chapter has strong links to psychoacoustical phenomena, such as the precedence effect, binaural loudness, and spaciousness. The chapter discusses which aspects of a concert hall give listeners the impression of intimacy and the perception of proximity to the sound. Moreover, it is explained how a concert hall can change the perceived dynamics of a music ensemble. Examples are presented using measured data from real concert halls.

1 Introduction

Concert halls are buildings dedicated to performing and listening to non-amplified music. Audiences gather to these venues to have the best possible acoustical conditions to enjoy live music, but also to socialize. The acoustical conditions of concert halls have been studied scientifically for more than a century, but the major part of the literature ignores important aspects of human spatial hearing. Traditionally, the halls are examined using impulse response measurements to collect objective data that can be compared between different concert halls. Most of these data are measured with omnidirectional microphones, thus ignoring important information that human spatial hearing benefits from. The reason for current approaches originate from the idea that these measurements give technically accurate and reliably reproducible results of the acoustical features. However, they do not describe accurately how concert halls are perceived by human listeners.

The most natural way to study music perception in concert halls is to listen to concerts in-situ and gather opinions from the audience as well as from the musicians

T. Lokki (✉) · J. Pätynen

Department of Computer Science, School of Science, Aalto University, Espoo, Finland
e-mail: tapio.lokki@aalto.fi

© Springer Nature Switzerland AG 2020

J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_7

173

and conductors. This method has been popular since Sabine (1900) published his fundamental work. Beranek (1962, 2012) has authored numerous articles and a few books, including comprehensive technical data, on the acoustics of concert halls. Moreover, seminal work in the area has been published by Hawkes and Douglas (1971), Barron (1988), and Kahle (1995). While in-situ listening with one's own ears is the most natural way to evaluate acoustics, the inherent problem is that the performed music typically varies from hall to hall. Unfortunately, human auditory memory hardly lasts 10 s (Sams et al. 1993), meaning that truly reliable comparison of the acoustics of halls is practically impossible between concert sites.

A major step to more detailed comparison was taken when binaural technology was adopted to room acoustics. Pioneering research was conducted in Germany by the groups in Göttingen (Schroeder et al. 1974) and Berlin (Kürer et al. 1969). They both involved dummy-head recordings, capturing the binaural sound in the studied concert halls. Also, both groups understood that the musical stimuli have to be the same in each hall to allow a valid comparison. To excite the hall with music, the Göttingen group used two omnidirectional loudspeakers on the stage, and they emitted an anechoic stereo recording of the 4th movement of Mozart's Jupiter Symphony. For laboratory listening tests, the binaural recordings were reproduced in an anechoic room using two loudspeakers with a cross-talk cancellation technique to preserve the binaural cues. In contrast, the Berlin group followed the Berlin Philharmonic Orchestra on their tour and recorded live music with a dummy-head, while the orchestra was playing the same music program in dress rehearsals in unoccupied halls. The listening tests were later performed in the laboratory using headphones, again preserving the binaural cues. Both of these teams found interesting results on the perceptual aspects of concert hall acoustics, but neither of these really concentrated on auditory spatial impressions or the benefits of spatial hearing.

The auditory spatial impression and other perceptual factors related to spatial hearing have been investigated in several studies. Marshall (1967) introduced the concept of spatial responsiveness as he proposed that narrow halls with high ceilings have more of such quality. In contrast, a wide hall with a low ceiling lacks spatial responsiveness. Based on these observations, it is clear that spatial responsiveness involves directional effects and, thus, binaural listening. In a review article, Marshall and Barron (2001) refer to the article by Kuhl (1978) (written in German), in which the connection between sound pressure level, lateral early reflections, and the degree of spatial impression is discussed. Keet (1968) showed earlier that an increased sound level produces an impression of spatial widening of the sound source. However, since then the level of the orchestral music has been mainly ignored in research. Instead, concert hall acoustics research has concentrated on the analysis of impulse responses. Some objective parameters of the ISO 3382-1 (2009) standard, such as j_{LF} , L_j or IACC can be applied to predict binaural properties of sound.

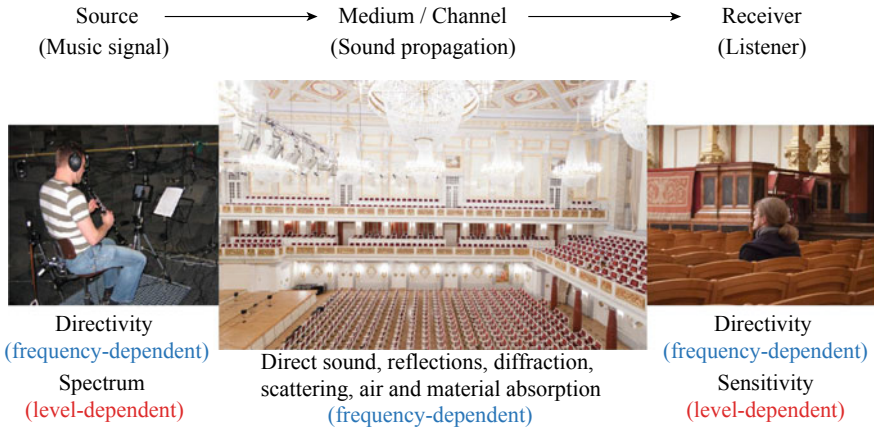


Fig. 1 The basic source—medium—receiver model in a concert hall

1.1 The Objective of This Chapter

The thorough review by Marshall and Barron (2001) describes the research in the 1900s on spatial impression in concert halls. This book chapter concentrates mainly on the research performed in the last decade. The main objective lies in explaining the room acoustical as well as psychoacoustical motivations, methods, and results of groundbreaking research in this field. Therefore, spatial hearing, psychoacoustics, and sound propagation in concert halls are discussed from a holistic viewpoint. Figure 1 presents the connection of musical instruments to human spatial hearing from the authors’ current perspective. In particular, to explain auditory spatial impressions, both the frequency and level-dependent aspects of music that propagates through a hall to listener’s ears have to be linked together. Pätynen et al. (2014) were among the first authors, who connected the well-known facts of dynamics-dependent spectra of orchestral instruments and the directional sensitivity of the human hearing to early reflections and their directions found in the room impulse responses. This connection has been further discussed by Lokki and Pätynen (2015) and Lokki (2016).

Most of the presented results are based on a state-of-the-art auralization system that allows authentic reproduction of concert halls in laboratory conditions. The auralization of the concert hall measurements are accomplished using the process illustrated in Fig. 2. The symphony orchestra is simulated on stage with 33 calibrated loudspeakers connected to 24 channels. The details of the loudspeaker orchestra setup can be found in previous publications of the authors (Lokki et al. 2011a, 2012; Pätynen 2011). The room impulse response from each of the loudspeaker channels is measured with a type 50-VI 3D vector intensity probe (G.R.A.S., Denmark) consisting of three co-centric phase-matched pairs of omnidirectional microphones capsules arranged on the x-, y-, and z-axes. The distance between the opposing capsules for

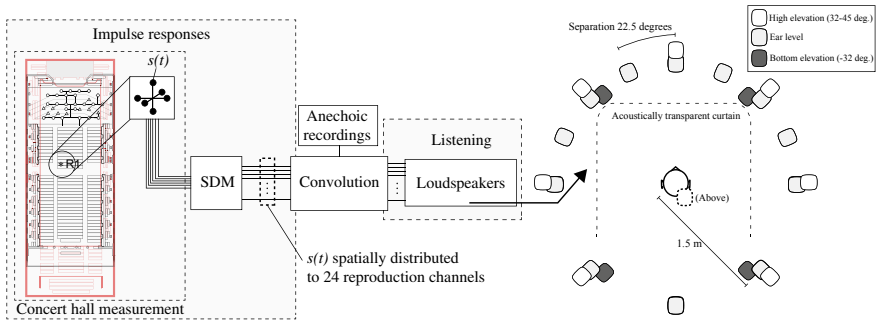


Fig. 2 The block diagram of the auralization scheme with loudspeaker orchestra measurement positions in the concert halls. The figure shows the method for a single source channel on the stage and the process is repeated for all sources to auralize the entire orchestra (Lokki et al. 2016)

each pair is 100 mm, and the impulse responses are measured with 48 kHz sampling rate using the logarithmic sine sweep technique (Farina 2000). The six impulse responses measured at a time are analyzed with the Spatial Decomposition Method (SDM; Tervo et al. 2013) that estimates the direction of incidence for each sample in an impulse response in short time windows. Based on the spatial information, the impulse response in the topmost omnidirectional microphone is distributed to reproduction loudspeakers as convolution reverberators. The distribution of samples is performed with the nearest loudspeaker technique in order to emphasize the spectral fidelity of the high frequencies (Pätynen et al. 2014) at the slight expense of spatial accuracy. Such a choice is adopted based on the earlier results, which clearly shows the importance of timbral fidelity over spatial fidelity (Rumsey et al. 2005). Finally, the anechoic recordings (Pätynen et al. 2008) are convolved with all reproduction channel responses. The distribution of the instruments to stage loudspeaker channels is the same method as described in Lokki et al. (2011a). The end result is a realistic reproduction of an orchestra in a concert hall, when the process is repeated for all sources on the stage.

2 Background and Motivation

Before the perceptual aspects can be discussed, there needs to be a discussion on the typical spatial room impulse responses measured in different concert halls. Two well-known concert halls in Berlin, Germany, namely the Konzerthaus and the Philharmonie serve here as examples. The former is an example of a classical rectangular hall, which is often referred to as “shoe-box” hall. The latter is a prime example of contemporary design in which the orchestra is located in the center of the hall, and the audience is surrounding the stage on multiple terraces or blocks, hence the moniker “vineyard” hall. Naturally, there exists also other general typologies, but these two

fundamentally different architectural examples highlight the differences in the spatial distribution of sound energy in a concert hall.

Figure 3 illustrates the measured cumulative sound energy distribution, averaged over 24 source positions on the stage, in the time-frequency-space domain (Pätynen et al. 2013). The analysis shows typical acoustical conditions of a shoe-box and a vineyard hall and highlights the differences across hall types. Although the illustrations here show only one seat in each hall, the other seats have similar properties in both halls. The bottom row shows the average cumulative frequency response at 5, 30, 200, and 2500 ms after the initial direct sounds and the spatiotemporal energy distributions use the same color coding for analyzed time windows.

Direct sounds and adjacent scattering, i.e., the initial 5 ms of the acoustic response arrives from each source on the stage in frontal directions. In a shoe-box hall, the stage floor is typically on the ear level of the audience at main parterre. Thus the listener does not receive the stage floor reflection, in contrast to halls with inclined seating areas. In Fig. 3b, it is seen that the seating rows behind the receiver position reflect sound within 5-ms time window. Figure 3e shows that there are indeed no seats behind the measurement position in the Philharmonie, as the response was measured on the last row of one audience block. The frequency responses illustrate that in the shoe-box hall, the direct sounds lack the low frequencies, but have considerably strong high frequencies. In contrast, in the vineyard hall with a raked audience area, the frequency response of the first 5 ms is quite different due to stage floor reflections.

Early reflections until 30 ms are visualized with dark blue color and they are integrated into the direct sound by the human auditory system. Two main differences between the example shoe-box and vineyard halls are the shape of the frequency responses and the spatial distributions of early sound energy. First, the shoe-box hall provides prominent lateral reflections inside the 30-ms time window, as illustrated by the triangular shape of the dark blue area in Fig. 3a. In addition, reflections from under the balconies are found in Fig. 3b, c. In the vineyard hall, the effect of the wall behind the measurement position can be clearly seen in Fig. 3d, e, f. In addition, the reflectors above the stage contribute to the cumulative energy. Nevertheless, there are hardly any side reflections, resulting in a distinct oval-shaped distribution of early energy in the lateral plane.

The second major difference lies in the frequency responses. As seen in Fig. 3g, the early reflections (between 5 and 30 ms) in the shoe-box hall strengthen the low frequencies below 200 Hz substantially, yet the middle frequencies up to 1 kHz remain at a relatively low level. The seat-dip effect causes such time-dependent filtering in halls with a flat floor and open seats (Tahvanainen et al. 2015). In contrast, when the sound cannot pass under the seats due to the chair construction and raked floor, the frequencies below 125 Hz are attenuated, and sound energy increases only slightly between 5 and 30 ms. At higher frequencies in this particular seat in the vineyard hall, the cumulative energy increases mainly due to the strong reflections from the wall behind the measurement position. It seems that when the reflection strengthens the direct sound from the stage floor, relatively weaker early reflections have little contribution to the cumulative sound energy after the strong direct sound. It could be argued that the direct sound accompanied by the floor reflection may mask the perception

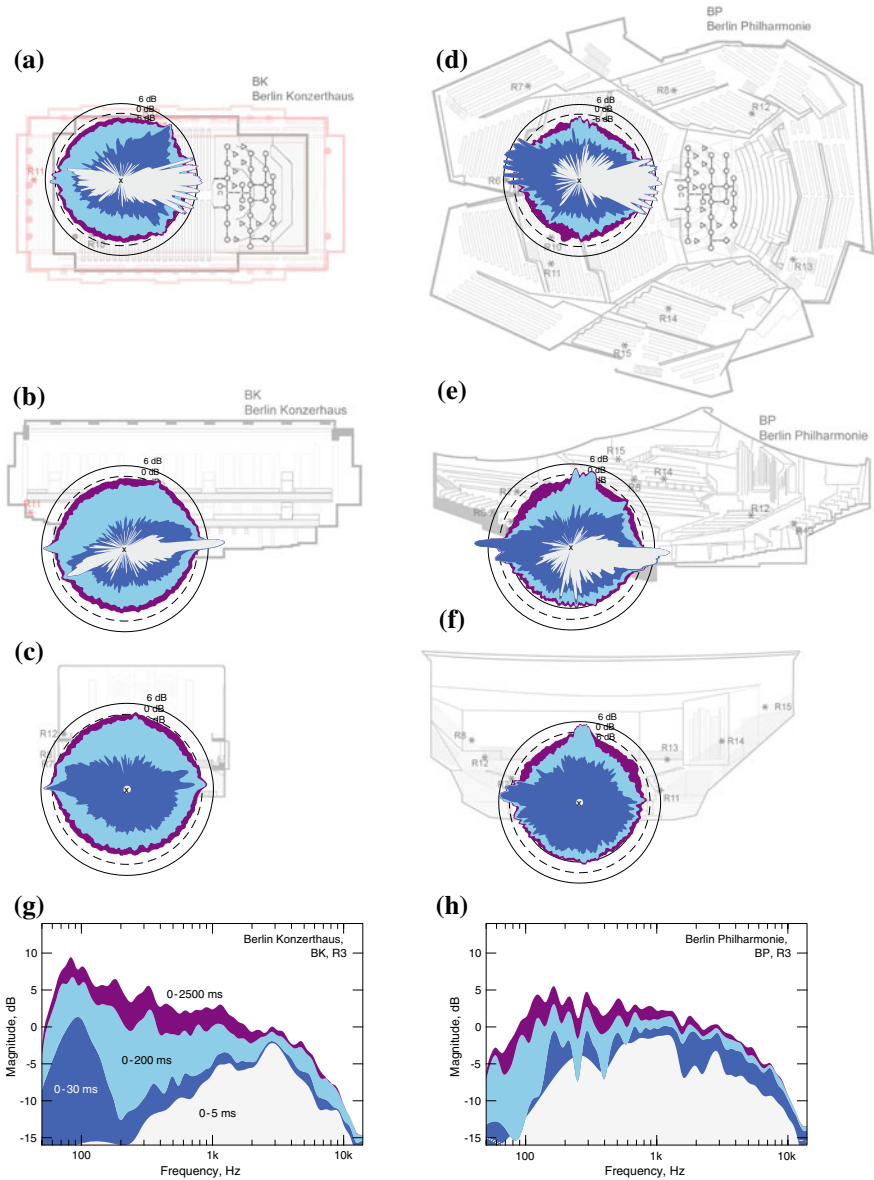


Fig. 3 Spatiotemporal and time-frequency analyses of Berlin Konzerthaus (left) and Berlin Philharmonie (right) concert halls at 15 m from the orchestra. Spatiotemporal visualizations are shown in lateral (panels ‘a’ and ‘d’), median (‘b’ and ‘e’), and transverse (‘c’ and ‘f’) planes in identical receiver positions. Bottom panels ‘g’–‘h’ visualize the temporal accumulation of the omnidirectional magnitude responses in respective positions. The time-windows of forward integration shown in panel ‘g’ are common for all plots

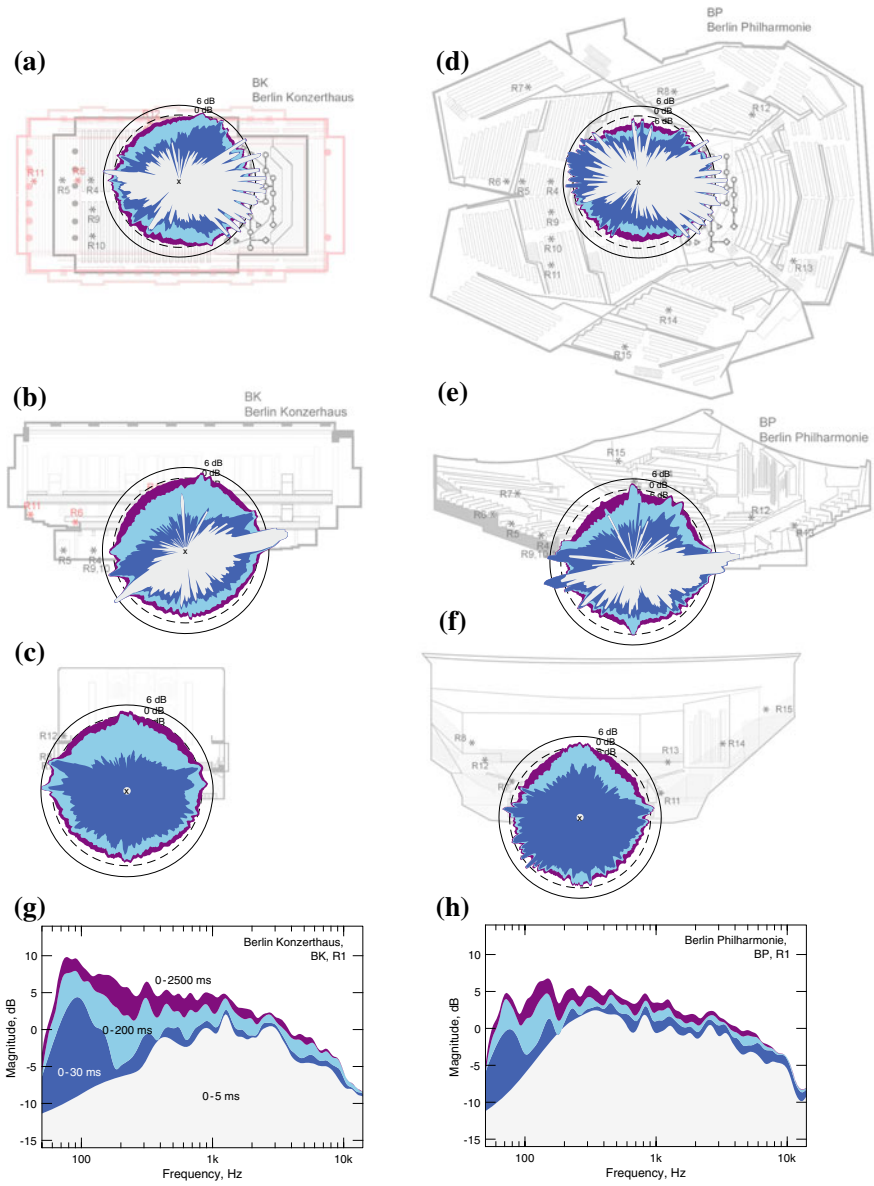


Fig. 4 Visualization of acoustic energy accumulation in two concert halls at a frontal receiver position at 7 m from the orchestra. The analyses are respective to Fig. 3

of early reflections. Marshall (1967, 1968) suggested similar ideas on masking, but hardly any research has been done on this complex perceptual phenomenon.

Later reflections between 30 and 200 ms increase the overall sound energy in both halls. In the shoe-box hall, the increase is particularly strong above 200 Hz, equalizing the frequency response to be more or less flat at 200 ms after the direct sound. Moreover, the energy in this time window reaches the measurement position almost evenly from all directions, and the incident energy in the measured positions has a round shape in all three visualization planes. In the vineyard hall, the energy distribution is not as uniform, and visualizations reveal an array of reflections from the ceiling and reflectors above the orchestra. Although the cumulative energy increases the overall level, the frequency responses retain their shape along the same peaks and dips as earlier. Sound arriving from the high elevation angle of the ceiling likely interferes with the subsequent floor reflection at the microphone array.

Finally, the **reverberation** beyond 200 ms increases the cumulative energy to its final state. In the Konzerthaus, the increase is spatially more uniform than in the Philharmonie. Notable differences between these halls can be observed in the smoothness of the overall frequency responses, level of low frequencies, and spatial distribution of sound energy. Figure 3d reveals one distinct late reflection from the right side of the measurement position. Such a reflection might be heard as an echo, or it might disturb orchestral balance by highlighting instruments in certain areas on the stage.

2.1 Auditory Impressions with a Real Orchestra in Example Halls

Among the most recent works on the perceived room acoustic quality, a study by Lachenmayr and Pätynen (2016) serves as a close counterpart to the early research by the Berlin group. In this particular study, the authors recorded the Staatskapelle Berlin at the dress rehearsals on the consecutive nights in the Konzerthaus and the Philharmonie, as the orchestra performed Beethoven's "Egmont Overture" in both halls. The recordings were captured with a four-channel pseudo-binaural arrangement with two channels on the sides of an absorbing sphere, and two additional rear surround channels with cardioid microphones. The sound was recorded simultaneously in two equidistant positions from the orchestra, resulting in four soundtracks to be compared. The reproduction utilized four loudspeakers of which two first were at $\pm 45^\circ$ angles, and the other two at $\pm 135^\circ$. The auditory evaluation took place in a relatively dry listening room with additional absorbing baffle immediately in front of the listener's head to reduce crosstalk across the front channels.

Although the technical approach was comparable to the preceding work of Schroeder et al. (1974) and Kürer et al. (1969), the authors adopted a different concept in the listening experiments. Instead of charting the acoustic quality in the traditional sense, the authors experimented with the potential variation in the

auditory impression during changes in music dynamics. As established earlier, music is not a static signal, but in reality, it contains continuous changes in many expressive aspects, such as instrumentation and dynamics. For this purpose, a 15-s excerpt with a gradual crescendo as the dynamic variation was isolated from the recordings for a listening test.

The listening test was a full paired comparison test with two repetitions, and the subjects had to choose the sample, which had more “impact.” This “impact” was defined as having more influence, being more interesting, or more effective on oneself. The subjects were asked to initially listen to both crescendos entirely instead of switching quickly back and forth. In addition to the paired comparison, the subjects were asked to write down one or more descriptive adjectives that described the perceived difference between the samples.

Two different listening tests were completed by 18 and 10 subjects, respectively. The first test reproduced the recordings as such with the possible loudness differences, and for the second test, the samples were loudness-matched. With the original stimuli, the result was clear; the subjects chose the Konzerthaus in both recorded seats having more “impact” than the Philharmonie. When the pairs were loudness-matched, listeners also reported greater impact in the Konzerthaus for the frontal seats but found no difference between the seats in the back of the two halls. However, the seat closer to the orchestra was always chosen over the farther seat, regardless of the hall.

The elicited descriptive adjectives reveal more detailed information on the perceived differences between the halls and the seats. There were three main differences related to proximity, loudness (strength+dynamics+crescendo), and spatial impression (envelopment+spaciousness+width). The results were in-line with the earlier research done with rudimentary simulations (Marshall and Barron 2001) and measured spatial impulse responses convolved with anechoic orchestral music (Pätynen and Lokki 2016a, b). In the following sections, these three aspects of concert hall acoustics are discussed in light of the recent research results. We also try to make links to more traditional psychoacoustics research in an effort to increase the common understanding on binaural human perception.

Before concentrating on the perceptual phenomena, the results of Lachenmayr and Pätynen (2016) are analyzed briefly in light of the objective measurements. The measurement positions, shown in Fig. 4, are close to the ones used in the recordings and are 8 m closer to the orchestra than measurement positions in Fig. 3. The main differences between the frequency responses (Fig. 4g, h) for different halls are found below 1 kHz. Here, the energy is accumulating quite differently as a function of time. In addition, in the Konzerthaus, there are 5 dB more low frequencies below 100 Hz. Figure 4c, f show clearly the spatial distribution of early sound energy. In the Konzerthaus, there are four lateral reflections from side walls and under the balconies. Moreover, the ceiling is quite high, resulting in later reflections from the ceiling than in the Philharmonie. The Philharmonie has also some lateral reflections, mainly on the left side, but also strong reflections from the reflectors above the orchestra. Thus, there is a clear difference in the spatial distribution of early energy. It is even possible that in the Philharmonie reflections in the median plane reach

the listening position earlier than the (relatively weak) lateral reflections. Marshall (1967) suggested that such order of reflections might result in less spatial impression, which was also the result in this listening test. Based on the recent formulation (Pätynen et al. 2014), the lateral early reflections in the Konzerthaus convey the high frequencies, emphasized in *fortissimo* playing, to the ear drums of a listener, as found in this listening test. In conclusion, when the orchestra makes a large *crescendo* from *pianissimo* to *fortissimo*, the largest differences in sound pressure level occur at low and high frequencies (Lokki 2016). Thus, the results by Lachenmayr and Pätynen (2016), obtained with the recordings of a real orchestra, are well supported by the time-frequency-space analysis of the measured impulse responses and the properties of human spatial hearing.

3 Early Reflections That Affect to Proximity, Intimacy and Engagement

One major purpose of music as an art form is to tell stories, evoke emotions, and touch the feelings of a listener. Therefore, it is not surprising that concert halls that sound intimate and engaging are often preferred. Over the years, researchers have called this aural aspect of a concert hall with different attributes, such as intimacy, proximity, presence, and engagement. As far as this is understood, they all address the same perceived phenomena, which are suggested to have a major positive influence on preference (Lokki et al. 2012; Kuusinen et al. 2014).

Intimacy is probably the most frequently used term (Beranek 1992). Beranek's description of intimacy characterizes the listening attribute as the closeness of communication between the listener and the orchestra. Moreover, Beranek (1992) identified an objective parameter, initial-time-delay-gap (ITDG), as "*the time between the arrival of the direct sound from the stage to the arrival of the first reflection at a measuring point*" for intimacy. However, the current understanding is that ITDG does not correspond well with intimacy, and ITDG has been misleading for many researchers (Hyde 2019). For example, consider a typical shoe-box hall in which at front rows the ITDG is much longer than in the last rows in the audience area. Everybody can easily understand that a frontal seat feels much more intimate than the seat in the back, although the ITDG suggests vice versa. In addition, ITDG ignores the overall level (i.e., perceived loudness) and spatial location of first reflections, proposing that both a ceiling reflection and a side wall reflection gives the same intimacy.

If ITDG does not explain intimacy, what is then the possible reason for a sound source to sound proximate? Lokki (2014) showed in their listening tests that the most intimate halls were preferred. Naturally, the sound pressure level is obvious as the louder the sound, the closer it is perceived. But perceived loudness does not explain everything; the loudest halls do not always sound the most proximate (Lokki 2014). The spatiotemporal analysis of measured impulse responses at the listening positions revealed that sound is perceived more proximate if there are strong lateral

reflections that reach the listener before the ceiling reflection. If the ceiling reflection, or sound from reflectors above the orchestra, is heard before lateral reflections, the sound is perceived more distant. The difference is seen in Fig. 3. In the Konzerthaus, the lateral and under-balcony reflections fit into the 30-ms window after the direct sounds, but in the Philharmonie, the situation is the opposite; the ceiling reflection is earlier than (weak) lateral reflections. It is possible that early lateral reflections reduce the interaural correlation, which could lead to the perception of a less distant sound, as suggested by Kendall (1995) for the relation of correlation and distance perception in stereo reproduction. Kuusinen et al. (2014) correlated many objective parameters with listening test data and found out that the lateral early energy fraction, j_{LF} , at high frequencies is associated with the perception of proximity. The j_{LF} is defined as the ratio of sound energies between 5–80 ms and 0–80 ms captured with figure-of-eight and omnidirectional microphones, respectively. This assumption is reasonable as research on spatial hearing has shown. Lateral reflections are louder than median plane reflections at the entrance of the ear canal because of directivity properties of the human head, as illustrated in Fig. 5.

Beranek (1992) wrote that lateral reflections are crucial for intimacy. He indicated repeatedly that intimacy in his terminology is the same perceptual phenomenon as spatial impression for Barron and Marshall (1981), who demonstrated the importance of early lateral reflections. Even though Barron and Marshall (1981) studied early lateral reflections with different sound pressure levels, they used the same recording with different levels missing the natural spectral change of music—see Sect. 4. Their methodology was, therefore, limited, which might be one of the reasons that in their studies the spatial impression was not affected by high frequencies over 1.5 kHz. This might have led to their simplistic conclusion that the most critical frequencies are the four octaves from 125 Hz to 1 kHz octave bands. For some reason, concert hall acoustics researchers still use only these mid frequency bands on many occasions. Meanwhile, Blauert and Lindemann (1986) and others have shown that spaciousness increases with increasing bandwidth of the later reflections. In fact, they concluded that all sound fields with components in the spectral range above 3 kHz produce a larger horizontal width than those which lack these components.

Furthermore, low frequencies have an important role for intimacy, and perception of strong bass is always connected to proximate sound (Lokki et al. 2012, 2016). Here, low frequencies should be analyzed down to 30 Hz, and not only in the 125 Hz octave band as usually done in concert hall acoustics research. The behavior of low frequencies is discussed in detail in Sect. 4.5. It is worth emphasizing that lateral reflections also render a stronger bass than ceiling reflections do, as seen in Figs. 3 and 4.

Finally, it has to be reiterated that intimacy is firmly a multimodal sensation. Hyde (2003) wrote an excellent report considering vision in addition to aural intimacy. As vision is the primary human sense, it could override aural intimacy in some cases. Furthermore, vision could also provide a baseline for intimacy. For example, when sitting in the first row of a balcony, visual cues suggest that orchestra is quite far away, and it is not expected that the music is really loud. Nevertheless, if music is loud, it might create the impression that this is a great hall as the orchestra can

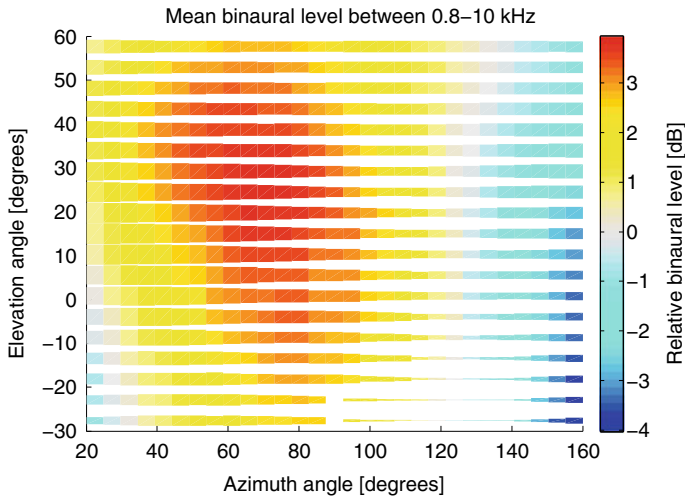


Fig. 5 Binaural level from regions on one side of the head at the frequency band of 0.8–10 kHz. The results show the characteristics of the binaural magnitude responses averaged over a region of $\pm 15^\circ$ in azimuth and elevation angles centered at the nominal angle in relation to the mean response at the frontal region ($\pm 10^\circ$ azimuth/ $0\text{--}30^\circ$ elevation). Thickest and thinnest bars denote a variation range of 0.7 and 16 dB, respectively

be heard so well at such a distance. Vice versa, when sitting parterre closer to the orchestra, it is expected to hear music at a certain sound pressure level. However, in many halls that lack early lateral reflections the sound is quite weak at frontal seats and the impression is far from intimate. The sensation is more like watching the orchestra playing in front, and the music is no longer optimally conveying emotions (Pätynen and Lokki 2015). It is evident that more audiovisual studies are needed to find out the multimodal perception of intimacy. Modern technology helps to bring both immersive visuals and audio to the laboratory, as done in a very recent study by Postma and Katz (2017). Interestingly, the authors found that subjects could be categorized into three subgroups; (i) subjects who judged auralizations more acoustically distant with increased visual distance, (ii) subjects who judged auralizations louder with increased visual distance, and (iii) subjects whose audio judgment was uninfluenced by the visual stimulus.

Kuusinen and Lokki (2015) also discuss the combination of visual and aural percepts for intimacy. Moreover, they focused on intimacy from an auditory and psychological perspective and viewed it as a dynamic feature, which is heavily influenced by the manner how musical expressions are translated and even enhanced by the acoustics of the hall. If a hall can provide dynamically varying spatial cues, which for instance can induce a perception of looming during *crescendos*, the experience of intimacy would be elevated not only by a heightened emotional response to the music, but also by a feeling of deeper involvement with the occupied space.

3.1 *The Quality of Early Reflections*

Robinson et al. (2013b) studied the effect of one side reflection on listeners' ability to separate two speech sources on stage with measured and simulated venues. The measurements were conducted in a 600-seat theater with a binaural headset worn by a male subject. For simulations, a simple concert-hall model with 11 distinct reflections and late reverberation were used, and in both cases, the binaural auralizations were presented via headphones. The listening tests were performed in two different laboratories and the task of the subjects was to indicate which one of the speakers, male or female, is on the left. In other words, the test investigated the spatial discrimination of multiple acoustic sources in real and simulated rooms, in which the properties of early reflections were modified.

The results of the listening tests were close to each other in both test conditions. With the real hall, the results show that the test participants could no longer distinguish which talker is on the left, when the proscenium splay surface—providing the first lateral reflection—was covered with a diffusor. In the case of a lateral reflection being even from a flat or from an absorptive surface, the task was easier. In other words, the experiment revealed that discriminating the lateral arrangement of two speech sources is possible at narrower separation angles when reflections come from flat or absorptive rather than diffusive surfaces. In the simulated hall, all 11 early reflections were rendered with measured or simulated diffusors. The results were similar to those for a real hall's results. It was easier to separate male and female speakers when the reflections were from the flat surfaces and diffusors hinder the subjects' ability to hear which one of the speakers was on the left (Robinson et al. 2013a).

The studies above were accomplished with speech stimuli, but it can be assumed that musical stimuli would have provided similar results. Diffusive architectural surfaces are applied widely in concert halls, but their perceptual consequences are not fully understood, and opinions in favor and against them exist (Oguchi et al. 2018; Kahle 2018; Marshall 2018). Robinson et al. (2013a, b) speculate that early diffusive reflections make it harder to localize the sources, suggesting that diffuse early reflections might blend sources better. According to Cremer and Müller (1982, p. 113), Meyer and Kuhl (1952) found that a sound source appears to perceptually expand laterally while still being localizable when placing large reflectors at both sides of the proscenium in the opera house in Hamburg. Unfortunately, the authors did not continue this line of research further.

Lokki et al. (2011b) investigated the perceptual consequences of the temporal envelope of the reflections. Commonly used diffusors in concert halls change the temporal envelope while reflections from a hard flat surface do not change the signals' phases. The waveforms and their temporal envelopes of a harmonic signal at auditory bands are illustrated in Fig. 6. Lokki et al. (2011b) suggested that with musical signals, as well as speech, the temporal envelope of reflections affect how well the precedence effect works. They proposed that reflections from diffusors might partially break down the precedence effect, resulting in less clarity. On the other hand, this also

means that sound sources might be perceived wider and less defined, which is often considered as a desirable sensation as the sound is then better blended. The effect of high-frequency scattering was also discussed by Kirkegaard and Gulsrud (2011). The authors reported that diffusers can produce a harsh sound. However, the sound quality improved when the diffusers were covered with absorptive material or flat panels.

3.2 Summary of Early Reflections

Early reflections are crucial for the quality of sound in a concert hall. Lateral reflections have been acknowledged for a long time to contribute positively to sound quality. However, often they are referred to as increasing the auditory source width, which is a misleading conclusion, as it is not always the desired property for a sound source. Instead, a proximate and engaging connection to music is wanted and early lateral reflections make sound louder and enable such connection.

Lateral reflections from flat or convex curved surfaces are integrated well to the direct sound in the human hearing system as long as they preserve the temporal envelope of signals. This is due to the precedence effect, which allows the auditory system to localize the first wavefront. It is important to note that the precedence effect does not render the early reflections inaudible. Early reflections increase the overall loudness, color the sound, and might change the perceived width of the source. As said, the temporal envelope preserving lateral reflections integrate to the direct sound best, increasing its quality and preserving the ability to localize. If such a reflection is coming from the median plane, i.e., from the ceiling or reflectors above an orchestra, the sound quality might be reduced due to coloration, which is the same in both ears. Moreover, such ceiling reflection might increase the interaural correlation, which could increase the perceived distance of the source (Kendall 1995).

If the reflections are scrambling the phases of upper harmonics, i.e., reflections from heavily diffusing surfaces, the precedence effect might partially break down, and such early reflections are not fully integrated with the direct sound (Lokki et al. 2011b). Such reflections might increase the perceived width of the source to the detriment of less defined location of the source. As a result, the instruments better blend together, but some listeners associate that to reduced clarity.

4 Time Varying Spectrum of Music

The previous sections illustrated how the room acoustics function as a transmission channel for the information expressed by the music signal. In particular, the role of early reflections was discussed. Although music is much more abstract than speech, music often aims to convey expressions or emotions. European-influenced classical music offers composers a variety of key elements for expressiveness, such as

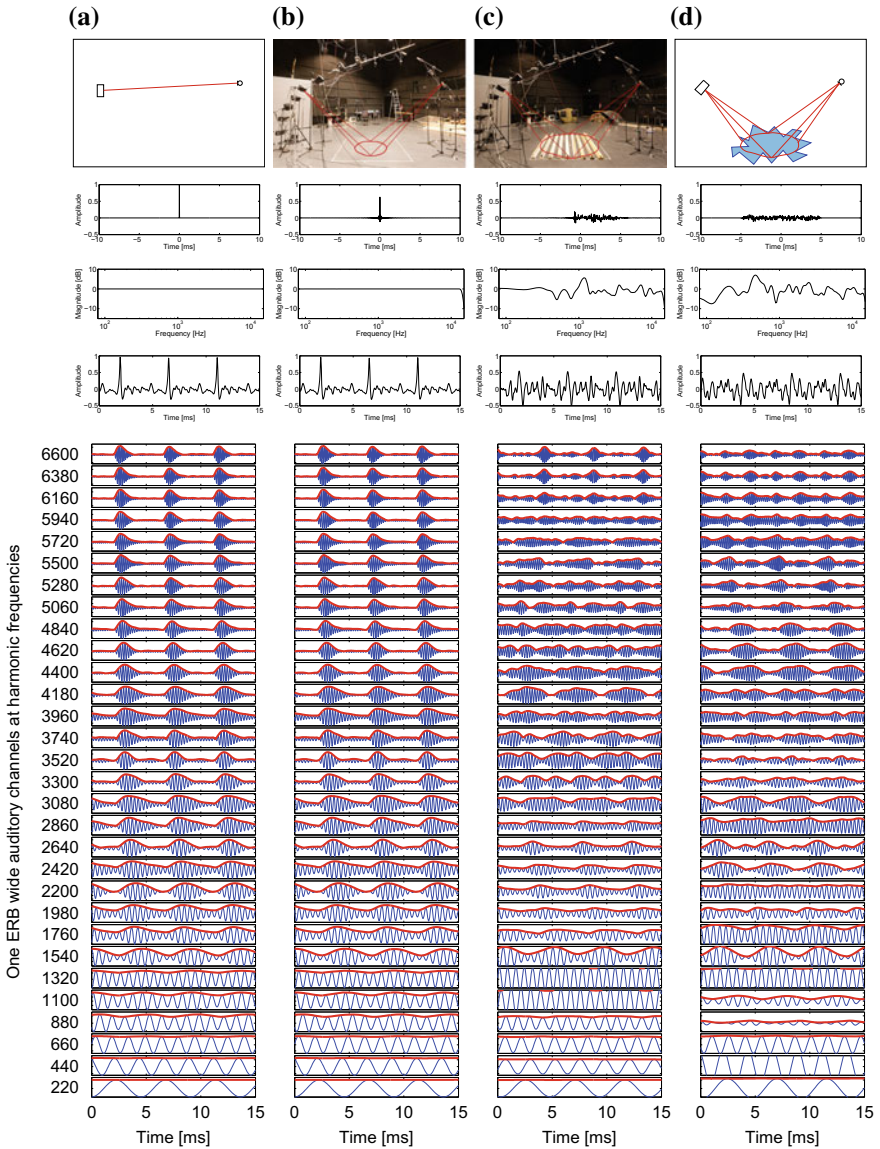


Fig. 6 From top to bottom: Illustration of a direct sound or a reflection, impulse response, frequency response, impulse response convolved with the trumpet sound of A_3 (220 Hz), amplitude and envelope of convolved trumpet sound at ERB wide bands, i.e., the waveforms (blue) and envelopes (red) at the outputs of auditory filters on the basilar membrane. **a** Direct sound **b** Temporal envelope preserving reflection **c** Temporal envelope destructing reflection at high frequencies **d** Temporal envelope destructing reflection at all frequencies

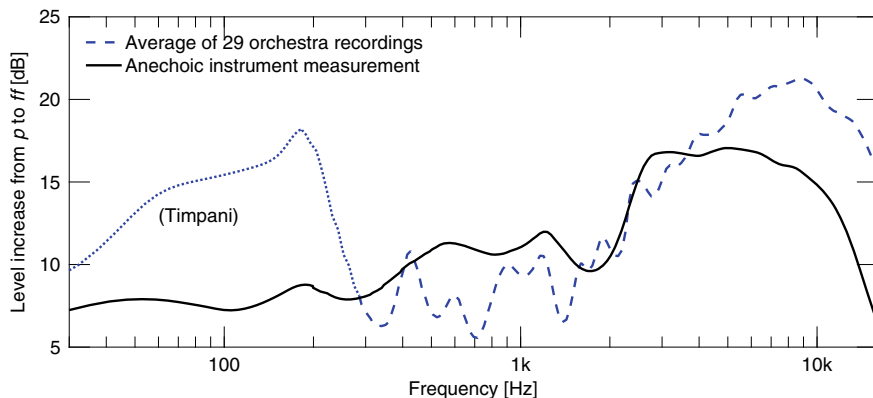


Fig. 7 Spectral change of full orchestra dynamics based on two datasets

pitch, note duration, timbre, and dynamics (Owen 2000, p. 6). Without the intended variations in the music dynamics and tone color, the expressiveness is often diminished. This, in turn, would impede the listeners' experience and the possible emotional impact sought from the performance. After all, experiencing the music is for many listeners the foremost reason for visiting concerts.

With this background, the importance of transmitted expressiveness becomes evident. Still, a survey on the related literature on room acoustics perception reveals that the music is mostly represented through quasi-stationary excitation signals. As a consequence, the time-varying properties related to the expressiveness—a key aspect of music—are typically not considered. Of the several aspects of expressiveness, recent research has concentrated on the music dynamics due to its relatively straightforward interpretation and simple measures. In order to include the aspect of music dynamics into the overall concept of perception of acoustics, the properties of orchestra instruments need to be discussed briefly.

The sound of most musical instruments is based on harmonic vibrations and pressure waves that are excited by the musician at a desired force and style. Depending on the type of the instrument, the amplitudes of harmonic overtones can vary strongly with the excitation. The higher harmonics are weakest at the minimum level of excitation, and the overtones become richer as the dynamics is increased. This effect applies to practically all instruments of a typical orchestra. The dynamic spectrum of individual instruments have been reviewed by Luce (1975), Meyer (2009), and others. These studies demonstrate that the most prominent dynamic spectrum effect is present with the brass instruments, where the magnitude slope of the overtone frequency envelope shows extreme variations between opposite playing dynamics. For many instruments, the amplitude of high-frequency overtones varies more than the amplitude of the fundamental. Consequently, the spectral content of the music signal varies disproportionately more at the high frequencies if the same pitches are played in different dynamic levels.

Two examples of orchestra dynamics are illustrated in Fig. 7. The diagram includes the spectral analyses of two datasets on orchestra instrument signals. The dotted line employs data collected from 29 commercially published recordings of Bruckner's Symphony No. 4, II. Movement, Bars 19–26. The particular passage includes a notable *crescendo* from indicated *pianissimo* to *fortissimo* with full orchestra. The excerpt is particularly useful for analyzing dynamics since the harmony and note pitches remain unchanged during the entire passage with only small variation in the rhythmic pattern. The played notes of the orchestral parts extend from B_2^b to F_6 , which corresponds to the approximate frequency range of 120–1400 Hz. The dashed curve in Fig. 7 shows the level difference between softest and loudest dynamic levels. As discussed above, the frequency range from 300 Hz to the highest fundamental of 1400 Hz is increased by 5–10 dB from *pp* to *ff*. In contrast, the frequency band consisting only of overtones gains up to 20 dB level increase during full-orchestra *crescendo*. The low-frequency emphasis is attributed to the strongly emphasized timpani tremolo near the *crescendo* peak.

The solid black line included in Fig. 7 presents the results from anechoic orchestra instrument measurements by Pätynen et al. (2008). In this dataset, separate notes of A-major triads were recorded with each instrument, spanning two octaves of the typical playing range of each instrument. All notes were recorded in indicated dynamics of *pianissimo* and *fortissimo*, and the maximum spectra of all notes were estimated from the signals. The number of each instrument in a typical orchestra complement of 83 players (without percussion instruments) was simulated in the overall spectrum. The result yields a similar trend as the first example, as the lowest frequencies up to 300 Hz show an average *pp*-to-*ff* level increase of 7 dB, middle frequencies up to 2 kHz gain around 10 dB, and the region of overtones increases by up to 16 dB.

While these examples illustrate the spectral effect of full-orchestra dynamics, they provide limited insight, considering the overall variation of dynamics in orchestral works. Contrasting dynamics with full orchestra occur relatively seldom, and more often, expressiveness is realized with variations in instrumentation and the texture of the instrument parts. For example, only some instruments or sections are performing during quiet parts, while other instruments, such as brass or percussions, join in for more powerful segments (Rimsky-Korsakov 1922). This effect strongly emphasizes the contrast between soft and loud passages. Therefore, it is feasible to analyze the distribution of frequency contents over a longer duration of music which contains a larger variety of instrumentation. For this purpose, the authors analyzed an orchestral recording of the entire first movement of Sibelius' Lemminkäinen suite. The recording was captured on a concert hall stage using closely positioned microphones for a commercial music production. The mix-down of the fairly dry signals of different sections give a representation of the orchestra sound over a 15-min piece.

The analysis in Fig. 8a presents the spectral variation as the distribution of occurrence for each frequency. In practice, the entire piece was segmented into one-second frames with 50% overlap, and the magnitude spectrum of each non-silent frame was stored. Naturally, each frame has a distinctive spectrum as different instruments overlap in each frame. Therefore, all magnitude values at each frequency bin (256 bins on a logarithmic scale) were ordered to obtain a rough distribution of spectral

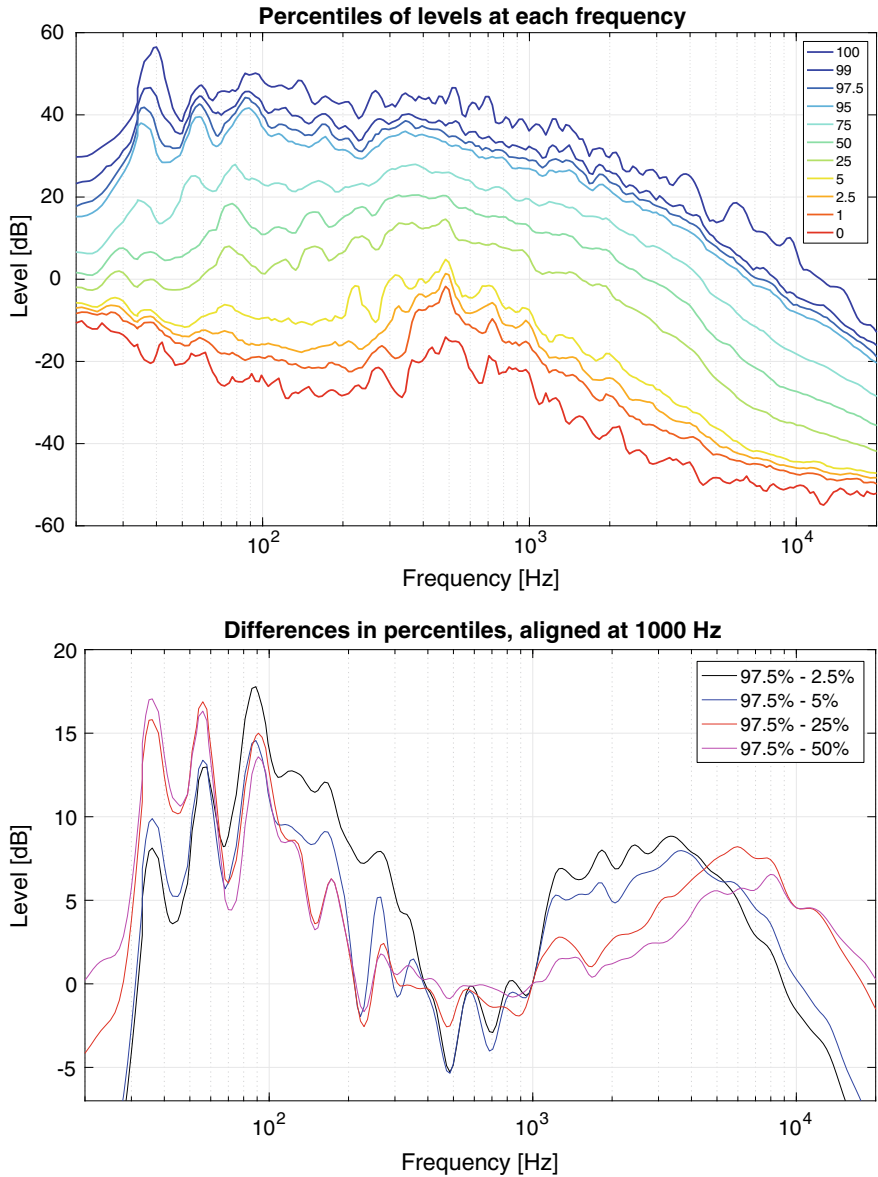


Fig. 8 **a** Distribution of an orchestra spectrum during a 15-min piece (Sibelius: Lemminkäinen suite, I movement) over one-second time frames as percentiles. **b** Differences between a representative *fortissimo* curve and various softer dynamics normalized to 1 kHz. The data are analyzed from a mix-down of a multi-channel near-field recording on a concert hall stage

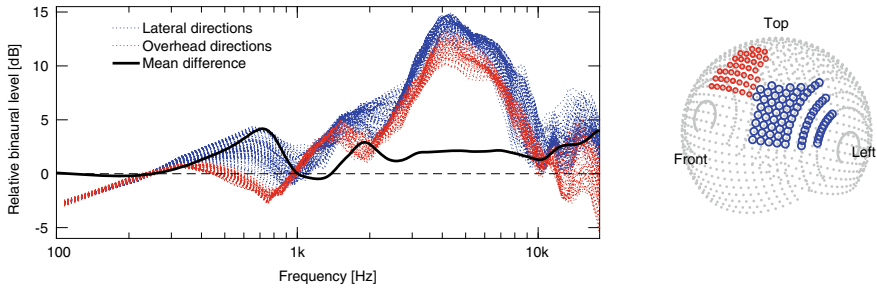


Fig. 9 Binaural levels for two typical directions for early reflections. Each magnitude response curve is an average over the CIPIC database of 45 HRTF measurements. The solid curve illustrates the average level advantage of lateral incidence (blue directions) to median plane reflections (red directions)

magnitudes at different frequencies over the entire piece. The resulting graph, Fig. 8a, shows the percentiles of each frequency bin, and thus the rough spectrum distribution over time. It can be theorized that the lowest 5-percentile represents a particularly quiet full orchestra *pianissimo*, or 25-percentile a typical soft *piano* played by part of the orchestra. By comparing these percentile curves to the full orchestra *fortissimo* curves of (97.5%), the analysis provides a statistical estimate for the dynamic spectrum change for the entire orchestra, including the typical orchestration. These estimates are found in Fig. 8b. The general trend shows a distinct similarity to the results shown in Fig. 7. In comparison to the 300–1000 Hz range, the frequency range of overtones is emphasized by up to 10 dB. Since this approach also takes into account the variation in music texture, the perceived low-frequency addition is thought to be much more prominent than in the investigations shown above. It should also be underlined that the largest level differences occur below 100 Hz, even down to 35 Hz.

Together the presented examples demonstrate that the spectral content of natural music signals is not constant, but it varies heavily with music dynamics. While the varying sound waves, regardless of the level and spectrum, are propagated linearly in rooms, the spectral sensitivity for level and direction of spatial hearing becomes an essential non-linear part of the entire source-medium-receiver model in Fig. 1.

4.1 The Dynamic Variation at High Frequencies Combined with Spatial Hearing

The acoustic effects of the head and outer ear are well-known through the research on binaural technology. One of the earliest reports of directional sensitivity dates back to the 1930’s (Wilska 2011), and the modern concept of head-related transfer functions (HRTF) has been developed through several studies overviewed by

Møller (1992) and others. Typically the mid- and high-frequency bands are perceived more sensitively with the ipsilateral ear when incident sound arrives from outside the frontal directions. Although the shadowing effect of the head can reduce the level at the opposite ear, the perceived effect with both ears' magnitude response combined is stronger for lateral incidence than with frontal sound. The magnitude of this effect is illustrated in Fig. 5, where the mid- and high-frequency binaural gains are the most consistent around azimuth angles of 40–80°.

The directional effect of spatial hearing plays a significant part when combined with the propagation paths provided by the room geometry and dynamic spectrum of the sound source. Given a frontal source, the room geometry yields the directions for the reflected sound, which is accumulated by the respective binaural magnitude response. Therefore, the reflection directions are instrumental in the binaural magnitude response of the room. Typical directions for early reflections in concert halls are the lateral angles from the side walls and the median plane directions from the ceiling or overhead reflectors. Detailed analysis of the binaural gain for such reflection directions in concert halls is depicted in Fig. 9. By comparing the average binaural gain between sets of lateral and median plane angles, it is evident that the lateral reflections yield emphasis on frequency regions of 400–1000 Hz and 1.7–10 kHz. Moreover, there are typically multiple lateral reflections, but usually only one ceiling reflection, as shown in Figs. 3 and 4.

The dynamic variation of the signal spectrum brings in the decisive component in the overall picture. Dynamic variation is emphasized at high frequencies, and lateral reflections, within angles of 40–80°, lead to binaural gain in the same high-frequency bands. Therefore, lateral reflections may increase the perception of dynamic variation. That was indeed proposed by Wettschureck (1976), who studied with speech the sensitivity of hearing for one reflection at 70 ms. The results showed that sensitivity of hearing is lower for late reflections from the side than those from behind or front of the listener when the listening level is high. At low listening levels, the sensitivities were more or less the same. Green and Kahle (2019) obtained similar results with music stimuli. Thus, it might be that audibility of reflections is a function of the listening level, and when the level is raised, the lateral reflections become more audible, increasing the perceived dynamics. However, more psychoacoustic research with real music at different dynamic levels is needed to understand level-dependent aspects of human spatial hearing better.

The earlier analyses of measured concert hall acoustics serve as visual examples of this concept. Figure 3a, d demonstrate how in certain halls the early response between 5 and 30 ms provide substantially more energy through reflections from the lateral angles. As illustrated in Fig. 9, lateral reflections emphasize frequencies in the 700 Hz to 1 kHz range, and to some extent also high frequencies—more so than frontal reflections due to the directional pinna distortions.

Another advantage in this respect lies in the second-order lateral reflections via the bottom surfaces of side balconies. This effect can be observed in the early energy along the transverse plane in Figs. 3c and 4c. When viewed from the listener position toward the stage, the conventional lateral reflections from the side walls are joined by an additional pair of reflections from moderately elevated directions. Such reflections

may also complement the overall timbre by providing additional early energy with slightly different HRTF spectra.

The given examples of concert halls also show that in some cases, the raked audience receives a direct sound which is amplified by the floor reflection (compare Fig. 3b, e). Correspondingly, distinct early reflections can be observed from ceiling, canopy, or reflectors (see Figs. 4e and 3e). In contrast to early lateral energy, reflections from the median plane may even reduce the binaural dynamic effect as the median plane incidence does not benefit from the directional emphasis of the dynamic-related frequency bands.

4.2 *Objective Metrics on the Dynamic Responsiveness with Spatial Hearing*

The proposed phenomenon of the sound field affecting the perceived music dynamics has been lately studied from objective and subjective viewpoints. Until recently, differences in dynamic effects in concert hall acoustics were only anecdotal references. For instance, Beranek has characterized this as the hall supporting both quiet and powerful dynamics: “*listening is enhanced immeasurably by the dynamic response of the concert hall*” (Beranek 2004, p. 509). Meyer, for one, has stated that the quality of *forte* is a sign of an acoustically excellent hall, while sound in quiet dynamics can also be acceptable in otherwise poorly rated halls (Meyer 2009, p. 199). Importantly, these remarks not only suggest the existence of a non-linear effect but also connect responsiveness of the hall to dynamics with subjective preference and increased listening pleasure.

Pätynen et al. (2014) explored the degree of the responsiveness to music dynamics by different concert hall acoustics in a study which combined the components of source, medium, and receiver as illustrated in Fig. 1. The dependency of the music signal spectrum was derived from anechoic orchestral-instrument measurements. The dynamic spectra of different instruments were mapped to the source positions respective to an orchestra layout in concert hall measurements. The spectra conveyed through the direct sound, and the early reflected sound to a binaural receiver were analyzed separately. The excitation of the left and right ears by the respective spectra were estimated with the model by Moore and Glasberg (1987), and the total binaural excitation was subsequently calculated with the binaural summation formula in the manner of Sivonen and Ellermeier (2006). In short, the adopted approach provides an objective metric for the auditory excitation by the orchestra sound in varying dynamic levels in different parts of the spatial room impulse response. This method was applied to ten European concert halls which were measured with the same calibrated system.

The main results revealed that the two hall typologies (rectangular shoe-box, or non-rectangular) varied prominently in the proportion of the binaural excitation between the direct sound and the early reflections in contrasting music dynamics,

hence the label “binaural dynamic responsiveness” (BDR). Expectedly, the effect was observed at the higher frequencies where the orchestra spectrum varies a lot between dynamic levels. When compared to the direct sound, the auditory excitation of the early reflections was greater in rectangular rooms than in non-rectangular rooms. In essence, this outcome explains the dynamic effect of concert hall acoustics. As discussed earlier, preceding studies have identified the perceptual effect of the spatial responsiveness of the acoustic space (Kuhl 1978; Marshall and Barron 2001). Those effects can also be easily connected to the recent results, as the rectangular rooms are often characterized by the early reflections in the lateral plane. The latter become proportionally more audible with increasing music dynamics. Further discussion on the regular features and their effect on the sound field related to music dynamics are presented by Pätynen and Lokki (2015) in an article which concentrates the inspection on few halls with different designs.

4.3 Perceptual Attributes of Music Dynamics Variation in Room Acoustics

While the objective approach has proposed means to quantify the dynamic responsiveness, controlled listening experiments have aimed to chart the perception of music dynamics in different acoustical conditions. A subjective listening test explored the perceptual attributes for different music dynamics in concert hall acoustics using a setting similar to the study with objective metrics. The presented music stimuli consisted of a short full-orchestral excerpt containing a sudden, yet musically feasible increase in the music dynamics while the instrumentation and orchestral texture were kept constant. The signal was created by concatenating Bars 41–43 (in piano) and 53–55 (in fortissimo) from an anechoic recording of Bruckner’s 8th symphony, II movement. Auralizations with room impulse responses from various concert hall measurements were presented to the listeners via a spatial 24-channel loudspeaker array in an acoustically treated listening space. The listening test employed a paired comparison method augmented by simultaneous free attribute elicitation. Hence, the subjects had first to decide which one of the presented two stimuli appeared to have a wider overall contrast between the different music dynamics. Additionally, the subjects gave short descriptions on which perceptual degrees the stimuli changed differently. Together these data provide insight on the overall perceptual dynamic responsiveness as well as the perceptual qualities of music dynamics in concert halls.

The results reported by Pätynen and Lokki (2016b) suggest that the foremost perceptual attribute differentiating the rooms’ responsiveness to music dynamics is the dynamic range itself. Out of circa one-thousand trials, approximately one fourth of the compared pairs demonstrated the dynamic range as the discriminating perceptual attribute reported by subjects. Another substantial attribute describing the effect of varied dynamics was the changing width of the auditory image. The

comparisons between six halls revealed that traditional room geometries (i.e., shoe-box halls) tend to provide a higher degree of perceived contrast between music dynamics than non-rectangular halls. This effect becomes more emphasized with increased receiver distances. These findings are consistent with earlier discoveries using objective metrics (Pätynen et al. 2014). Among individual rooms, Vienna Musikverein and Berlin Konzerthaus appeared to exhibit particularly strong dynamic loudness and spatial effects among the concert halls included in the experiment (Pätynen and Lokki 2016b). Correlation analysis, to describe salient connections between perceived dynamic responsiveness and traditional objective room-acoustic parameters, showed that the strength (G) at high frequencies and the inverse of early binaural coherence [$1 - \text{IACC}$] predicted the high degree of dynamic effects best. Contrary to expectations, the early lateral energy fraction did not show substantial correlation with the perceptual effect.

4.4 Dynamic Responsiveness and Emotional Impact of Listening

Earlier sections in this chapter presented the communication of dynamic variations as one ingredient for the emotional impact in music listening. Another study by Pätynen and Lokki (2016a) assumed a more general perspective on the concept of music dynamics. While the study described above aimed to explore the perception of music dynamics, the second set of experiments focused on the emotional impact produced by listening to orchestral music for different acoustical conditions. The employed listening test methodology departed from conventional experiments by applying psychophysiological measurements during focused listening session with participants. A multi-channel spatial sound reproduction system in a laboratory setting proved to be a feasible environment to measure of electrodermal activity, i.e., variations of skin conductance due to autonomic nervous system activation. The subjects were presented a sequence of 12 auralizations of a total of six concert halls via convolutions of anechoic orchestra material and spatial room impulse responses. The music signal was a positively looming passage from Beethoven's 7th symphony, first movement. Bars 11–18 of the piece begin softly with alternating woodwind chords and ascending major scales with strings, and eventually culminates in a prominent full-orchestra crescendo to the tonic A-major fortissimo. With its easily approachable tonal development and texture, and as one of the principal works in the orchestral literature, the passage was regarded a suitable excerpt for comparing the possible emotional effects between orchestra performances. The duration of each auralization was approximately 30s, and the excerpts were presented in randomized order with 15-s silence in between without any listener interaction.

During the entire experiment of circa 12 min, the Skin Conductance Response (SCR) was recorded synchronously with the presented audio stimuli. Following the conventions for analyzing this kind of measurement, the intensity of emotional

responses in different acoustic conditions could be ranked according to the recorded psychophysiological data. Similarly to the study on perceived music dynamics, the acoustic characteristics of rectangular halls showed a distinct advantage also in eliciting stronger emotional responses. Of individual rooms, the two highest mean SCR responses were found in the front positions at Vienna Musikverein and Berlin Konzerthaus. Listening to the performance in a non-rectangular hall instead of a shoebox room appeared to have a negative effect on the emotional intensity comparable to the approximate doubling of the listening position's distance from the orchestra.

In order to gather more evidence for the emotional impact measured through psychophysiological responses, the set of 12 auralizations were also presented to the same subjects in a more traditional listening experiment with paired comparisons (Pätynen and Lokki 2016a). Listeners were asked to choose the more impactful stimulus, and these self-reported results show the same general pattern as the SCR results.

To summarize, spatial hearing combined with non-linear spectral excitation of musical instruments and different acoustic propagation paths and directions due to concert hall designs yields a complex setting. The communication of music dynamics elicits a wide range of perceptual attributes linked to the constantly varying nature of music. Therefore, the perception of room acoustics does not remain static for any signal, but instead, the music itself influences it. With the recent experiments on these topics, previously presented claims and impressions have found support from the research findings.

4.5 Dynamic Variations at Low Frequencies Combined with the Seat-Dip Effect

Figure 8a illustrates a considerable amount of energy at low frequencies, even below 40 Hz, in symphonic music. In addition, large dynamic changes are the strongest at low frequencies (see Fig. 8b), often due to the orchestration as presented earlier. Therefore, it is reasonable to briefly discuss the low frequencies in concert halls, although, at low frequencies, the spatial hearing does not play any role for auditory impression. Nevertheless, Marshall and Barron (2001) mention that the perceived width of the sound source is wider if the music is loud and the bass is strong.

As the sound travels from the stage over the seats at near grazing angles below 15° at low frequencies an excess attenuation has been measured by Sessler and West (1964), and Schultz and Watters (1964) already 50 years ago. The phenomenon is called as the seat-dip effect, and it is observable for the direct sound and some early reflections. The seat-dip effect is a combination of several phenomena, but mainly results from diffractions from the seat rests and from floor reflections that interfere destructively with the direct sound (Ishida 1995). Bradley (1991) suggested that the main frequency of the attenuation depends on the dimensions of the seats. Tahvanainen et al. (2015) confirmed Bradley's results by analyzing measured data

from 10 different concert halls, and they also found that the inclination of the floor and the seat type affect the range of attenuated frequencies. Based on these measurements, the seat-dip effect can be categorized into two cases, wide-band attenuation centered around 150–300 Hz, and narrow-band attenuation centered around 100 Hz (Tahvanainen et al. 2015). These two cases are indeed seen in Fig. 4g, h. In the Konzerthaus, with flat audience floor and lightweight open seats, the main attenuation occurs at frequencies of about 200 Hz, and the attenuation spans up to 1 kHz. However, the frequency response is filled in by the later reflected energy so that the attenuation due to the seat-dip effect has disappeared already 200 ms after the direct sound. In contrast, the Philharmonie, with a raked audience area and seats that do now allow the underpass of sound, has a narrow seat-dip around 100 Hz. This is not corrected at all with the later reflected energy.

Together with the early reflected energy, the seat-dip in these two halls most probably contributes to the frequency response below 100 Hz as well. In both halls, it can be seen as a positive interference at about one octave lower than the seat-dip frequency. In the Konzerthaus, the seat-dip frequency is one octave higher than in the Philharmonie, and therefore this positive interference is also at higher frequencies. Also, the emphasis on frequencies below 100 Hz is much stronger in the Konzerthaus, and we assume that this boost is related to the seat-dip effect. Such low-frequency behavior is very important for the dynamic variation in music, as illustrated in Fig. 8b, the largest dynamic differences are between 35 and 100 Hz. Therefore, it is reasonable to assume that in the Konzerthaus the dynamics at low frequencies are larger than in the Philharmonie (Lokki and Pätynen 2020).

5 Late Reverberation Contributes to Loudness, Envelopment, Spaciousness, and Timbre

As discussed in Sect. 3, it is clear that early reflections have a crucial role in the perceived acoustics of a concert hall. Indeed, Haapaniemi and Lokki (2014) found that the characteristics of a hall are recognized within the first 80 ms of an impulse response. They investigated measured real concert halls with multichannel auralization system so that while the late reverberation was the same in each rendering, the first 80 ms was from different halls. Subjects had to choose out of four different renderings, which hall was the same as a reference, and the recognition rate was close to 100%. If the first 80 ms was kept constant (i.e., from one hall) and the late reverberation tails were from different halls, the subjects considered the task much harder, however the recognition rate was still about 80%. The main attributes that subjects used in recognizing the halls were timbre and auditory width.

Regardless of the big role of early reflections, the characteristics of late reverberation, i.e., length of decay, spectral coloration, and spatial distribution, are very important for the quality of music. Indeed, reverberation ties music segments together and blends instruments, making music more enjoyable. When people are

comparing different concert halls, they always pay attention to reverberance and aspects that the reverberation causes, such as envelopment, loudness, and width to some extent (Lokki et al. 2016). Reverberation also colors the sound, and, in particular, the strength of high frequencies in reverberation influences to perceived brightness, dynamics, and brilliance (Pätynen and Lokki 2015). Moreover, listeners often perceive a certain “warmth” if there are enough low frequencies (down to 20 Hz) and a reasonably long low-frequency decay.

From the binaural point of view, the major perceptual effect of late reverberation is related to envelopment, i.e., how well a listener feels surrounded by music. Again, in the authors’ studies (e.g., Lokki et al. 2016), it was found that the hall type has a great influence on the envelopment. In shoe-box halls with flat-floor audience areas, there is typically good envelopment, and late reverberation is uniformly distributed around the listener, see Figs. 3 and 4. In contrast, in vineyard halls with raked audience areas, the envelopment is reduced, and late reverberation is less uniformly spread. The perceptual impression is often that in the halls with highly raked audience areas the listeners are “looking at the music” (Pätynen and Lokki 2015) while in the flat floor halls the listeners feel “inside the music”.

There is not much research on the importance of the perceptual viewpoint on the spatial distribution of the late reverberation. Recently, Lachenmayr et al. (2016) studied the directional effect of reverberation for the perceived envelopment. The results were not very clear, but they propose that when reverberation arriving from the side or above is reduced, the feeling of envelopment decreases. The frontal or rear reverberation had minor, although important effects on envelopment. In a second listening test, the listeners adjusted one late component of the sound field blindly to the preference level. The results showed that when the hall lacks reverberant energy from a certain direction, subjects raised energy at that particular direction to a level so that reverberation is more or less uniform in the end. For example, in the Philharmonie, the lacking reverberation behind the listener was compensated more than in the Konzerthaus.

Another contribution to the late directional reverberation has been presented by Kahle (2016). The author concluded that excessive reflections and reverberation from frontal directions can have a negative influence on orchestral balance and on on-stage hearing conditions for musicians. Kahle (2016) also describes several halls in which the openness and quality of sound were increased when the back wall of the stage and choir balcony was covered with absorptive material. Thus reducing the level of frontal reflections and reverberation increased the quality of sound both on the stage and in the audience area.

6 Conclusions

Concert halls are often studied by measuring impulse responses and computing room acoustical parameters based on the measurements. By definition, the impulse responses are linear and time-invariant, and the spatial aspects of the sound field could

be investigated using directional microphones or a dummy head as a measurement device. However, they do not reveal anything about the level-dependent phenomena, such as source broadening according to level or the effect of room acoustics to the perceived dynamics of music.

This chapter describes in detail, why the traditional means of analyzing concert hall acoustics are insufficient. It is explained how the level dependent spectra of musical instruments and sensitivity of human hearing are important to understand to room acoustics. In particular, the phenomena related to spatial hearing and perception of music in concert halls are explained.

Acknowledgements This research was supported by the Academy of Finland, project nos. 296393 and 289300. The authors thank two anonymous reviewers for constructive comments.

References

- Barron, M. 1988. Subjective study of British symphony concert halls. *Acta Acustica united with Acustica* 66 (1): 1–14.
- Barron, M., and A. Marshall. 1981. Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. *Journal of Sound and Vibration* 77 (2): 211–232.
- Beranek, L.L. 1962. *Music, Acoustics, and Architecture*. New York, NY, USA: Wiley.
- Beranek, L.L. 1992. Concert hall acoustics—1992. *The Journal of the Acoustical Society of America* 92 (1): 1–39.
- Beranek, L.L. 2004. *Concert Halls and Opera Houses: Music, Acoustics, and Architecture*, 2nd ed. New York, USA: Springer.
- Beranek, L.L. 2012. *Concert Halls and Opera Houses: Music, Acoustics, and Architecture*. Berlin: Springer Science & Business Media.
- Blauert, J., and W. Lindemann. 1986. Auditory spaciousness: Some further psychoacoustic analyses. *Journal of the Acoustical Society of America* 80 (2): 533–542.
- Bradley, J. 1991. Some further investigations of the seat dip effect. *Journal of the Acoustical Society of America* 90 (1): 324–333.
- Cremer, L., and H. Müller. 1982. *Principles and Applications of Room Acoustics*. London, England: Applied Science Publishers.
- Farina, A. 2000. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proceedings of the 108th Audio Engineering Society (AES) Convention*, Paris, France, preprint no. 5093.
- Green, E., and E. Kahle. 2019. Dynamic spatial responsiveness in concert halls. *MDPI Acoustics* 1 (3): 549–560.
- Haapaniemi, A., and T. Lokki. 2014. Identifying concert halls from source presence vs room presence. *Journal of the Acoustical Society of America* 135 (6): EL311–EL317.
- Hawkes, R.J., and H. Douglas. 1971. Subjective acoustic experience in concert auditoria. *Acta Acustica united with Acustica* 24 (5): 235–250.
- Hyde, J. 2003. Acoustical intimacy in concert halls: Does visual input affect the aural experience? (multisensory integration and the concert experience). Pub. P.S. Veneklasen Research Foundation: Santa Monica, MA, USA.
- Hyde, J.R. 2019. Discussion of the relation between initial time delay gap (ITDG) and acoustical intimacy: Leo Beranek’s final thoughts on the subject, documented by Jerald R. Hyde. *MDPI Acoustics* 1 (3): 561–569.

- Ishida, K. 1995. Investigation of the fundamental mechanism of the seat-dip effect—Using measurements on a parallel barrier scale model. *Journal of the Acoustical Society of Japan (E)* 16 (2): 105–114.
- ISO 3382-1. 2009. Acoustics – measurement of room acoustic parameters—part 1: Performance spaces. International Standards Organization.
- Kahle, E. 1995. Validation d'un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras. PhD thesis, Université du Maine.
- Kahle, E. 2016. Acoustic feedback for performers on stage—return from experience. In *International Symposium on Musical and Room Acoustics (ISMRA)*, La Plata, Buenos Aires, Argentina.
- Kahle, E. 2018. Halls without qualities—or the effect of acoustic diffusion. In *The 10th International Conference On Auditorium Acoustics*, 169–173, Hamburg, Germany.
- Keet, W.V. 1968. The influence of early lateral reflections on the spatial impression. In *Proceedings of the 6th International Congress on Acoustics*, vol. 3, E53–E56, Tokyo, Japan.
- Kendall, G.S. 1995. The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal* 19 (4): 71–87.
- Kirkegaard, L., and T. Gulsrud. 2011. In search of a new paradigm: How do our parameters and measurement techniques constrain approaches to concert hall design? *Acoustics Today* 7 (1): 7–14.
- Kuhl, W. 1978. Räumlichkeit als Komponente des Raumeindrucks [Spaciousness as a component of spatial impression]. *Acta Acustica united with Acustica* 40 (3): 167–181.
- Kürer, R., G. Plenge, and H. Wilkens. 1969. Correct spatial sound perception rendered by a special 2-channel recording method. In *The 37th Audio Engineering Society Convention*.
- Kuusinen, A., and T. Lokki. 2015. Auditory distance perception in concert halls and the origins of acoustic intimacy. In *The 9th International Conference on Auditorium Acoustics*, 151–158, Paris, France.
- Kuusinen, A., J. Pätynen, S. Tervo, and T. Lokki. 2014. Relationships between preference ratings, sensory profiles, and acoustical measurements in concert halls. *Journal of the Acoustical Society of America* 135 (1): 239–250.
- Lokki, T. 2016. Why is it so hard to design a concert hall with excellent acoustics? In *The Second Australasian Acoustical Societies' Conference (Acoustics 2016)*, Brisbane, Australia, invited Plenary lecture.
- Lokki, T. 2014. Tasting music like wine: Sensory evaluation of concert halls. *Physics Today* 67 (1): 27–32.
- Lachenmayr, W., and J. Pätynen. 2016. Influence of acoustics on emotional impact of music in Konzerthaus and Philharmonie Berlin. In *DAGA 2016, 42, Jahrestagung für Akustik*, 79, Aachen, Germany.
- Lachenmayr, W., A. Haapaniemi, and T. Lokki. 2016. Direction of late reverberation and envelopment in two reproduced Berlin concert halls. In *The AES 140th International Convention*, 9503, Paris, France.
- Lokki, T., and J. Pätynen. 2020. Objective analysis of the dynamic responsiveness of concert halls. *Acoustical Science and Technology*, 41 (1): 253–259.
- Lokki, T., and J. Pätynen. 2015. The acoustics of a concert hall as a linear problem. *Europhysics News* 46 (1): 13–17.
- Lokki, T., J. Pätynen, A. Kuusinen, H. Vertanen, and S. Tervo. 2011a. Concert hall acoustics assessment with individually elicited attributes. *Journal of the Acoustical Society of America* 130 (2): 835–849.
- Lokki, T., J. Pätynen, S. Tervo, S. Siltanen, and L. Savioja. 2011b. Engaging concert hall acoustics is made up of temporal envelope preserving reflections. *Journal of the Acoustical Society of America* 129 (6): EL223–EL228.
- Lokki, T., J. Pätynen, A. Kuusinen, and S. Tervo. 2012. Disentangling preference ratings of concert hall acoustics using subjective sensory profiles. *Journal of the Acoustical Society of America* 132 (5): 3148–3161.

- Lokki, T., J. Pätynen, A. Kuusinen, and S. Tervo. 2016. Concert hall acoustics: Repertoire, listening position and individual taste of the listeners influence the qualitative attributes and preferences. *Journal of the Acoustical Society of America* 140 (1): 551–562.
- Luce, D.A. 1975. Dynamic spectrum changes of orchestral instruments. In *Audio Engineering Society Convention 51*, paper No: 1025.
- Marshall, A.H. 1967. A note on the importance of room cross-section in concert halls. *Journal of Sound and Vibration* 5 (1): 100–112.
- Marshall, A.H. 1968. Levels of reflection masking in concert halls. *Journal of Sound and Vibration* 7 (1): 116–118.
- Marshall, A.H. 2018. On the architectural implications of diffusing surfaces. In *The 10th International Conference On Auditorium Acoustics*, 618–624, Hamburg, Germany.
- Marshall, A.H., and M. Barron. 2001. Spatial responsiveness in concert halls and the origins of spatial impression. *Applied Acoustics* 62 (2): 91–108.
- Meyer, J. 2009. *Acoustics and the Performance of Music*. New York, NY, USA: Springer.
- Meyer, E., and W. Kuhl. 1952. Bemerkungen zur geometrischen Raumakustik [Comments on geometric room acoustics]. *Acta Acustica united with Acustica* 2 (2): 77–83.
- Møller, H. 1992. Fundamentals of binaural technology. *Applied Acoustics* 36 (3): 171–218.
- Moore, B.C.J., and B.R. Glasberg. 1987. Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns. *Hearing Research* 28: 209–225.
- Oguchi, K., M. Quiquerez, and Y. Toyota. 2018. Acoustical design of Elbphilharmonie. In *The 10th International Conference On Auditorium Acoustics*, 89–96, Hamburg, Germany.
- Owen, H. 2000. *Music Theory Resource Book*. New York, NY, USA: Oxford University Press.
- Pätynen, J. 2011. A virtual symphony orchestra for studies on concert hall acoustics. Ph.D. thesis, Aalto University School of Science.
- Pätynen, J., and T. Lokki. 2015. The acoustics of vineyard halls, is it so great after all? *Acoustics Australia* 43 (1): 33–39.
- Pätynen, J., and T. Lokki. 2016a. Concert halls with strong and lateral sound increase the emotional impact of orchestra music. *Journal of the Acoustical Society of America* 139 (3): 1214–1224.
- Pätynen, J., and T. Lokki. 2016b. Perception of music dynamics in concert halls. *Journal of the Acoustical Society of America* 140 (5): 3787–3798.
- Pätynen, J., V. Pulkki, and T. Lokki. 2008. Anechoic recording system for symphony orchestra. *Acta Acustica united with Acustica* 94 (6): 856–865.
- Pätynen, J., S. Tervo, and T. Lokki. 2013. Analysis of concert hall acoustics via visualizations of time-frequency and spatiotemporal responses. *Journal of the Acoustical Society of America* 133 (2): 842–857.
- Pätynen, J., S. Tervo, P.W. Robinson, and T. Lokki. 2014. Concert halls with strong lateral reflections enhance musical dynamics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 111 (12): 4409–4414.
- Pätynen, J., S. Tervo, and T. Lokki. 2014. Amplitude panning decreases spectral brightness with concert hall auralizations. In *Proceedings of the 55th Audio Engineering Society Conference*, Helsinki, Finland. New York, NY, USA: Audio Engineering Society, paper No: 49.
- Postma, B.N.J., and B.F.G. Katz. 2017. The influence of visual distance on the room-acoustic experience of auralizations. *Journal of the Acoustical Society of America* 142 (5): 3035–3046.
- Rimsky-Korsakov, N. 1922. *Principles of Orchestration (tr. E. Agate)*. Berlin: Editions Russes de Musique.
- Robinson, P., J. Pätynen, and T. Lokki. 2013a. The effect of diffuse reflections on spatial discrimination in a simulated concert hall. *Journal of the Acoustical Society of America* 133 (5): EL370–EL376.
- Robinson, P., J. Pätynen, T. Lokki, H.S. Jang, J.Y. Jeon, and N. Xiang. 2013b. The role of diffusive architectural surfaces on auditory spatial discrimination in performance venues. *Journal of the Acoustical Society of America* 133 (6): 3940–3950.

- Rumsey, F., S. Zielinski, R. Kassier, and S. Bech. 2005. On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *Journal of the Acoustical Society of America* 118 (2): 968–976.
- Sabine, W. 1900. Reverberation: Introduction. *The American Architect*.
- Sams, M., R. Hari, J. Rif, and J. Knuutila. 1993. The human auditory sensory memory trace persists about 10 sec: Neuromagnetic evidence. *Journal of Cognitive Neuroscience* 5 (3): 363–370.
- Schroeder, M.R., D. Gottlob, and K.F. Siebrasse. 1974. Comparative study of european concert halls: correlation of subjective preference with geometric and acoustic parameters. *The Journal of the Acoustical Society of America* 56 (4): 1195–1201.
- Schultz, T., and B. Watters. 1964. Propagation of sound across audience seating. *Journal of the Acoustical Society of America* 36 (5): 885–896.
- Sessler, G., and J. West. 1964. Sound transmission over theatre seats. *Journal of the Acoustical Society of America* 36 (9): 1725–1732.
- Sivonen, V.P., and W. Ellermeier. 2006. Directional loudness in an anechoic sound field, head-related transfer functions, and binaural summation. *Journal of the Acoustical Society of America* 119 (5): 2965–2980.
- Tahvanainen, H., J. Pätynen, and T. Lokki. 2015. Analysis of the seat-dip effect in twelve European concert halls. *Acta Acustica united with Acustica* 101 (4): 731–742.
- Tervo, S., J. Pätynen, A. Kuusinen, and T. Lokki. 2013. Spatial decomposition method for room impulse responses. *Journal of the Audio Engineering Society* 61 (1/2): 16–27.
- Wettschureck, R. 1976. Über die Abhängigkeit raumakustischer Wahrnehmungen von der Lautstärke [On the dependence of room acoustic perception by sound level]. Ph.D. thesis, TU Berlin, Germany.
- Wilska, A. 2011. Untersuchungen über das Richtungshören [Studies on directional hearing]. Technical report, Aalto University.

Auditory Room Learning and Adaptation to Sound Reflections



Bernhard U. Seeber and Samuel Clapp

Abstract Sound reflections are abundant in everyday listening spaces, but they are rarely bothersome, and people are often not even aware of their presence. As shown in several studies, this is partially due to adaptation of the human auditory system to the spatiotemporal reflection pattern, namely, through an increase in the echo threshold that follows repeated exposure to the same reflection pattern. This raises the question of whether adaptation mechanisms to room reflections lead to improved localization accuracy as well—a measure more tangible for everyday listening. Moreover, this benefit would only be useful if it could be maintained through changes in the reflection pattern such as those produced by head turns or body movement within the room, or from sources at different locations. Therefore, a particular mechanism is hypothesized by the current authors based on learning a representation of the room geometry, rather than learning of or adapting to a specific reflection pattern. This chapter reviews and discusses the available literature on the build-up of the precedence effect and related effects in speech understanding in reverberation. In light of the hypothesis of room learning, it aims to trigger a discussion about the underlying mechanisms.

1 Introduction

One of the most remarkable abilities of the human auditory system is how it can function successfully in highly challenging acoustic environments. Nearly every built environment—where humans spend most of their time—contains surfaces that reflect acoustic energy. When a sound is emitted in such a space, listeners not only receive a signal that is traveling directly from the sound source to the ears but multiple

B. U. Seeber (✉) · S. Clapp
Audio Information Processing, Department of Electrical and Computer Engineering,
Technical University of Munich, Arcisstrasse 21, 80333 Munich, Germany
e-mail: seeber@tum.de

S. Clapp
e-mail: samuelclapp@fb.com

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_8

delayed copies of the original signal superimposed upon it. The specific pattern of these delays (in time and space) is determined by the geometry of the surfaces in the environment and the positions and orientations of both the sound sources and the listeners.

It would be extremely difficult to localize sound effectively in reverberant environments without any sort of mechanism to deal with reflected sound energy. However, as most know from personal experience, normal-hearing human listeners are quite good at this, even in the absence of other sensory cues (Hartmann 1983; Blauert 1997). The mechanisms underlying this ability are understood in the context of the so-called “Precedence effect”, a name given to a group of related phenomena that are briefly discussed in the following—compare Blauert (1997), Litovsky et al. (1999), Brown et al. (2015). The precedence effect has been studied extensively with stimuli played from both a single leading location and a single lagging location, separated by some time interval on the order of milliseconds. Results from one listener in such an experiment by Seeber and Hafter (2011) are depicted in Fig. 1. For short lead-lag delays up to 12 ms (here for a spoken word), a single sound location was reported both in the localization responses (top) and in the fusion responses (bottom). Hence, despite being played from loudspeakers separated by 60°, at short delays the lead and lag stimuli are perceptually fused into a single auditory event. This (fused) auditory event is located between the loudspeakers for delays up to 2 ms, an effect known as “summing localization”, which is widely used in stereophony. For longer delays the sound is perceived as coming from the leading loudspeaker, hence the name “precedence effect” or “localization dominance” of the first incoming wavefront. Above the “echo threshold”, here around 12–24 ms, lead and lag are segregated into two distinct auditory events, one perceived at the lead and one at the lag location. “Lag discrimination suppression” is the third phenomenon besides localization dominance and fusion subsumed under the term “precedence effect” (Yang and Grantham 1997; Litovsky et al. 1999; Brown et al. 2015). It indicates the listener’s difficulty to determine binaural parameters of the lag stimulus at short lead-lag delays, with the difficulty decreasing as the delay increases.

In such precedence effect experiments, stimuli are usually presented in isolation, whereas in most listening situations, sources repeatedly emit sound, thereby giving the auditory system the opportunity to reassess and integrate information about the source and the room over time (Clifton and Freyman 1996; Hafter 1996). The focus in the current chapter is on the role of the context, that is, on the question of how signals that immediately precede the test stimulus affect the perception of that stimulus. This is an important question in terms of understanding spatial hearing in rooms because, in the course of a normal day, people spend minutes or hours at a time in one place, and repeatedly in the same places from day to day. This gives our auditory system the chance to collect information about the space via the acoustic signals reaching the ears. There is much evidence, both in studies of the precedence effect and in the articles discussed in this chapter, that the accumulation of this acoustic information results in later reflections being suppressed in favor of the direct sound, shown by significant increases in the echo threshold and lag discrimination suppres-

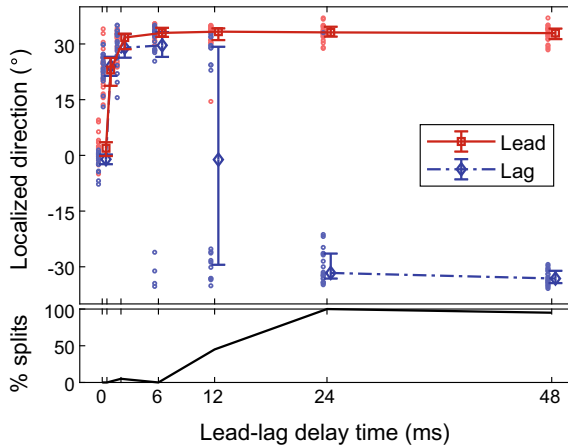


Fig. 1 (Top) Localized positions in the precedence-effect paradigm. The same sound, a recording of the spoken word *shape*, is played at equal level from both the lead loudspeaker (plotted at $+30^\circ$) and the lag loudspeaker (at -30°), separated by the lead-lag delay given on the abscissa. Plotted are individual responses (**small circles**) as well as medians with quartiles. Medians are connected by lines for readability. Results were obtained with a light-pointer method. Lead and lag locations were randomized (from $\pm 30^\circ$) and the listener was in one run instructed to point to the leftmost sound, if two locations were perceived, and in another run to point to the rightmost sound location. The listener thus pointed in separate trials to the lead and lag locations. By knowledge of the actual lead and lag location, data were analyzed into pointing to the lead (**red squares**) and the lag (**blue diamonds**). **(Bottom)** Percentage of trials in which more than one sound event, that is, split images, was perceived. All data stem from one normal-hearing listener. Replotted from Seeber and Hafer (2011)

sion. This process is assumed to assist with localization and speech understanding in reverberation.

One way to understand this phenomenon is to consider it as inhibition and adaptation process. The auditory system, after having obtained sufficient information about the acoustic environment, suppresses information from the directions of strong reflections. This view is grounded in discrimination suppression experiments that show reduced access to binaural cues in the lagging sound. Starting with early views of the precedence effect as inhibition (McFadden 1973), corresponding models use inhibition of monaural and binaural information after the sound onset as a general suppression process or suppress specific lead-lag delays or interaural time differences (ITDs), that is, time differences in the arrival of a signal at the closer ear and the arrival at the farther ear—compare Hartung and Trahiotis (2001), Lindemann (1986a). For example, in such models, inhibition equipped with a forgetting time increases upon repeated presentation from the same direction, thus demonstrating adaptation in terms of an increasing echo threshold. A model proposed by Djelani and Blauert (2002) exhibits a direction-specific buildup of the precedence effect that could be viewed as an adaptation of binaural neurons that signal particular directions, namely, the echo threshold is increased for directions from which reflections were

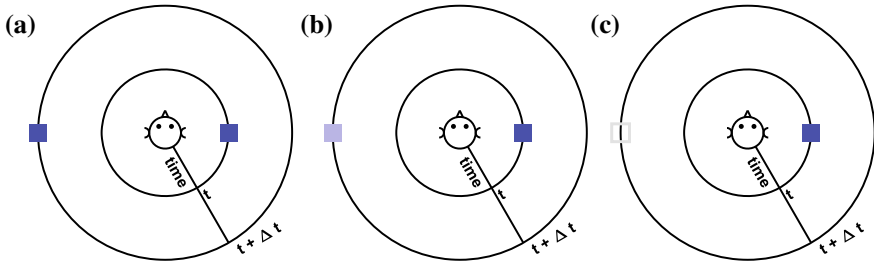


Fig. 2 A schematic diagram of the buildup of the precedence effect. In these plots, the radius represents time, and the azimuth the direction of the sound stimulus. The sub-figures represent the stages of the build-up of the precedence effect for the case of a leading click coming from the right of the listener, followed by a lagging click from the left, separated in time by some interval Δt . **(a)** Initial exposure to the lead-lag pair, where both clicks are perceived. **(b)** Intermediate phase after a few presentations of the lead-lag click pairs, where the lagging click is still perceived but is beginning to be suppressed. **(c)** Complete build-up of the precedence effect after repeated presentation of lead-lag click pairs, where the lag is still acoustically present but no longer perceived as a discrete auditory event

repeatedly presented. Here the question comes up of whether such direction-specific adaptation is ecologically useful, or even ecologically valid, given that real-world acoustic environments are much more complex. Figure 2 shows spatiotemporal diagrams that illustrate direction-specific adaptation.

The (direction-specific) adaptation observed in precedence effect experiments and addressed in the present chapter should be considered separate from “binaural adaptation”, a term coined by Hafer for phenomena related to the localization of binaural click trains that do not contain reflections (Hafer 1996). Hafer and colleagues studied the relative weight given to individual clicks in a click train when localizing the complete click train. For high-rate click trains, localization is determined almost exclusively by the first click, indicating onset dominance (Hafer and Dye 1983; Stecker and Hafer 2002). A restart of the adapted binaural system, seen by an increased weight of a click, occurs, for example, after a gap in the train (Hafer and Buell 1990).

The common way of acoustically experiencing a room is not a static process since listeners and sound sources almost never remain completely fixed in place. In addition, listeners constantly make small adjustments to their head orientation and posture. This strongly affects the interaural cues. Thus, listeners usually perceive a specific space by experiencing its reflection pattern that varies in time. In a simple “adaptation” process, any movement of sources or listeners would thus require an ongoing re-adaptation to the current configuration. Of course, as long as the listener remains within the same room, these changes in the reflection pattern will abide by an underlying logic as dictated by the geometry of the room—a natural scenario that listeners encounter daily. The current authors thus postulate another potential way to understand this process, namely, through “room learning” or “abstraction” as proposed in Seeber et al. (2016), Menzer and Seeber (2014). Rather than sim-

ply suppressing information impinging on the listeners from specific directions as in an adaptation process, in the “abstraction” process, the direct sound and early reflections are used by the auditory system to develop an abstract representation of the room geometry. The positions of reflecting surfaces (and thus the room boundaries) are inferred via the timings and locations of reflections. Based on such an abstract room representation, signals arising from early reflections can be anticipated and, subsequently, their interaural binaural cues can be suppressed. While this appears indistinguishable to an adaptation process for static sources, a room abstraction process can allow for suppression even after head turns or position changes in the room. Thus, in such a process, the perception of a given room would not require a new period of adaptation following source or listener movements, once the auditory system has acquired the necessary information to develop a model of the geometry of the space. A schematic diagram of the two proposed processes following head rotation is depicted in Fig. 3. In this chapter, several studies will be examined in light of these two proposed processes—adaptation and abstraction—with the aim of exploring whether the results provide evidence of the existence of one or the other, or even of both.

2 Context Effects with Simple Lead-Lag Stimuli

In seeking to understand how the auditory system deals with reflections, many studies have made use of the simplest case, with a leading stimulus from one horizontal location and a lagging stimulus from another horizontal location, separated by a time interval. In the *real world*, this would correspond to a room with a single reflective surface, and all other surfaces being completely absorptive. The direct sound would come from the lead location and the reflection from the lag location. In the introduction, it was discussed how these lead-lag stimuli are perceived based on the duration of the interval between the lead and the lag. Here we consider how stimuli that immediately precede a test stimulus affect the perception of the latter. This is particularly important with respect to everyday listening in rooms, where longer periods of time are spent in specific spaces and thus, the acoustic context plays an important role.

One of the first and best-known examples for the impact and build-up of context is the “Clifton effect” (Clifton 1987). A lead-lag click pattern with an interstimulus interval of 5 ms was played in free field from loudspeakers located at $\pm 90^\circ$, that is, from perpendicularly to the left and the right of the listener. First, a click train of several seconds was played, with the right loudspeaker leading. Then, the positions of the lead and lag were suddenly flipped, and listeners were asked whether they heard clicks from the left loudspeaker, right loudspeaker, or from both. Immediately after the locations were switched, all listeners heard both the leading and the lagging clicks distinctly. However, after a certain number of clicks in the new spatial orientation, listeners returned to hearing clicks solely in the direction of the lead loudspeaker.

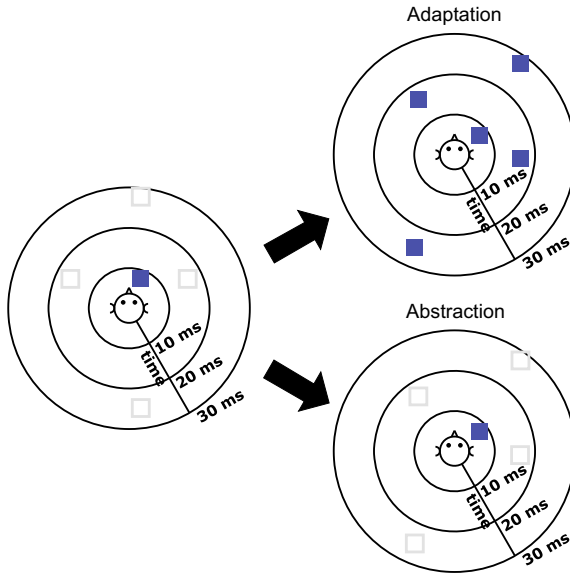


Fig. 3 A schematic diagram illustrating predicted perception in a room (simplified to the direct sound and four first-order reflections) following head-turning under the “adaptation” and “abstraction” processes. The left scheme represents perception after repeated exposure to the direct sound with reflections, where only the direct sound is perceived distinctly and the reflections are suppressed. Both schemes on the right show the new orientation of the direct sound and reflections with respect to the listener following the listener turning the head 30° to the left. The expectation in an “adaptation” process (**top right**) is that the precedence effect breaks down and needs to build back up again because the head-turning results in a new spatial orientation of reflections with respect to the listener. In an “abstraction” process (**bottom right**), however, the listener continues to suppress the reflections in favor of the direct sound because the room is still the same

This experiment demonstrates the involvement of dynamic processes in the precedence effect, with both fusion and the echo threshold increasing over time in response to repeated exposure to the same stimulus. The Clifton effect was modeled by Djelani and Blauert (2002) based on an approach by Lindemann (1986a, b) by using an interaural cross-correlation function with a dynamic inhibition algorithm. Peaks in the cross-correlation function (corresponding to the directions of sound events) inhibited the function at other delay values (i.e. other directions). In addition, the strength of the inhibition was increased when it was triggered regularly and repeatedly. The model successfully reproduces the results of Clifton (1987). At the first presentation of a lead-lag stimulus, two peaks in the binaural activity map appear, corresponding to both the lead and lag locations and indicating a situation above the echo threshold, where fusion has not yet taken place. However, after 3–4 presentations of the lead-lag stimulus at regular intervals, the peak corresponding to the lag has disappeared, indicating that fusion has now taken place. The Lindemann model assumes an adaptation process for replicating the Clifton effect and does not need to estimate the room geometry, just the spatial locations of lead and lag sound sources as inferred

from the binaural signals. The binaural features of the lag become suppressed over time. Consequently, the lag auditory events cease to be spatially separate.

The psychoacoustic data by Rachel Keen (then Clifton) indicate that suppression of the lag does not happen immediately. She examined different click pair rates and found that at a rate of 1/s, the time required for lag suppression to build up completely was 8–10 s, while for faster click pair rates of (2–4)/s, the time was only 3–5 s. Thus, it appears to require 8–12 click pairs for the lag to be suppressed rather than a set length of exposure time to a given reflection pattern (Clifton and Freyman 1989, 1996). Freyman et al. (1991) confirmed the total number of click pairs in the conditioning train to be the most salient quantity in predicting an increase in the echo threshold, rather than the click pair rate or the total duration. An increase in the echo threshold was observed when increasing the total number of click pairs from 3 to 5 to 9, with only a very small change when increasing from 9 to 17 clicks. This suggests that a plateau is reached with nine click pairs in the conditioning train. Note that while click trains are an interesting stimulus since they provide a quantified amount of information per click, build-up appears to occur faster for continuous speech (Djelani and Blauert 2001).

It is interesting that it requires a number of click pairs (greater than one) to increase the echo threshold, as a new click pair in a train does not actually contribute any new information, as it is identical to the preceding ones. This is congruent with an adaptation process, where an inhibitory process requires repeated stimulation to build up over time. Likewise, it could also be explained by the room learning process, which requires repeated presentation in order to extract reliable information, namely, a sufficient number of observations.

In a controlled laboratory setting, the click pairs in the conditioning train and in the test stimulus can be made exactly identical. However, it would be extremely rare for this type of scenario to happen in a natural environment, so for these effects to have any sort of validity outside of the laboratory, it is of interest to know whether they can be observed for stimuli that have the same spatiotemporal arrangement, but differ in other aspects.

In another experiment by Freyman et al. (1991), clicks were replaced with short white noise bursts, using either the same or different noise tokens for every burst. The echo threshold increased in both cases, indicating that the exact waveform of the bursts is not critical. Clifton et al. (1994) went on to vary the frequency content and intensity of the test click pair with respect to the conditioning train, but now they measured the lag discrimination ability rather than the echo thresholds. There was little to no difference in discrimination performance when changes in frequency content or intensity were introduced between the conditioning click pairs and the test click pair, versus when these parameters were held identical between the two pairs. These results indicate that the spatiotemporal arrangement of the click pairs is the salient feature that goes with increased echo thresholds and lag discrimination suppression.

The results derived from different noise tokens and intensity changes between conditioning train and test stimulus can be understood via either the adaptation or the abstraction process. However, it is unclear whether a tonotopically operat-

ing model would predict the psychoacoustic results of cases where the frequency content differs between conditioning train and test stimulus, as might happen in a real environment with a time-varying signal. The model of Djelani and Blauert (2002) passes the ear signals through a band-pass filter bank before activating the inhibition algorithm in its binaural processor. Thus, the build-up of inhibition in one critical band should not necessarily cause build-up in another one, unless that band receives the same reflection pattern. To fix this problem, binaural filters could be made much wider than peripheral auditory filters. In contrast, a model employing abstraction does not have a problem explaining this result. Room geometry is not frequency-dependent, and so a given spatiotemporal reflection pattern holds for any stimulus, regardless of its frequency content. Therefore, if the auditory system can build an internal geometric model of a room, it does not matter if there is a difference in the frequency content of the conditioning and test stimuli.

Another interesting question is whether conditioning trains that consist of clicks at only the lead or only the lag location can affect the echo threshold. In an abstraction process, one would not expect a single click to have an effect on echo thresholds, as a single click implies an anechoic environment with no room geometry information. In an adaptation process, it is, however, possible that a single click might reduce the sensitivity to other locations. This holds when inhibition would be expected to build up at locations other than the conditioning click location, albeit the Clifton effect shows direction specificity in the build-up.

Freyman et al. (1991) found that when only the lead or only the lag click pairs were presented in the conditioning train, it resulted in a reduced echo threshold for the test click pair, as compared to the case where no conditioning train was used. Thus, listeners were more attuned to reflections after hearing clicks at just the lead or just the lag location in the conditioning train. Freyman and Keen (2006) confirmed these findings and showed that the echo threshold even reduces to that of the single click baseline. In their experiment 3, the build-up click train was followed by a lead-only click train. Echo thresholds were reduced, but not to that without the build-up click train, indicating a partial break-down. When only a single lead- or lag-only click was inserted into the end of the build-up click train, echo thresholds were unaffected—a certain number of lead-only or lag-only clicks seems needed to disturb the build-up. A recent study by Bishop et al. (2014) also looked at the effects of lead- and lag-only clicks in the conditioning train. After the conditioning train, a 4-s test stimulus of click pairs (with time delays varying across trials) was played, and listeners were asked how many clicks at the lag location they heard. When a lead-alone conditioning train was used, approximately 9% more lag clicks were heard as compared to the case of a silent conditioning stimulus, whereas a lag-alone conditioning train resulted in approximately 7% fewer lag clicks being heard (averaged across all listeners and lead-lag time delay values). Thus, in this study, conditioning clicks at the lead location slightly increased the sensitivity to the lag location, in agreement with Freyman et al. (1991). However, the conditioning clicks at the lag location slightly decreased sensitivity to the lag location, in contrast to the results of Freyman et al. (1991). None of these results directly support either the

adaptation or the abstraction hypothesis, but the absence of strong effects without a lag click pair being present agrees with the abstraction hypothesis.

One other question that might be asked is the following. Once an echo threshold has been increased for a specific lead-lag orientation, how long does it persist in periods of silence? In an abstraction process, it could persist indefinitely, as a period of silence would not tell the auditory system that the listener has moved to a different space unless one assumes a forgetting time constant for the surrounding room. Similarly, in an adaptation process, it would depend on the forgetting time constant of the inhibition algorithm. Keen and Freyman (2009) looked at the effect on echo thresholds of a test click pair when the conditioning click pair train was followed by a variable amount of silence. They found virtually no difference in echo threshold after up to 3 s of silence following the conditioning train compared to when the test click was presented immediately after the conditioning train. This finding could support either process and in the future longer periods of silence could be investigated to determine a value for the forgetting-time constant.

The underlying mechanisms of the precedence effect and its build-up with continuous stimuli are more difficult to ascertain. Various studies have shown that onsets and offsets are weighted more heavily, as these periods in time are thought to give the most unimpaired information about the locations of the direct sound and reflections (e.g., Houtgast and Aoki 1994; Stecker and Hafter 2002). This onset dominance can be used, for example, to improve spatial coding with cochlear implants (Monaghan and Seeber 2016). However, it is also known that localization dominance is caused by the temporally overlapping part of continuous noises (Dizon and Colburn 2006; Seeber 2011), suggesting that, generally, information from temporal modulations is used. How room abstraction and adaptation processes would function for ongoing stimuli is difficult to judge without further study. Generally, identifying the locations of individual reflections from ongoing stimuli either for spatially specific suppression of binaural cues of reflections in an adaptation process or for an abstraction of room dimensions from individual reflections remains an issue which is not yet well understood.

3 Break-Down

So far, the building up of room adaptation or abstraction through repeated exposure to a given reflection pattern has been discussed, and it has been shown that, once it has been built up, it can persist for several seconds. Several studies have investigated the reverse process, namely, whether the build-up state breaks down when introducing new stimuli after an initial conditioning train. During a typical day, a listener will move from space to space, and each new space will require an adjustment in terms of which directions are suppressed. This raises the question of how long echo suppression persists for the *previous* space. Does it disappear immediately, or does it persist for some time following exposure to a new pattern?

Djelani and Blauert (2001) investigated this idea by varying the spatial orientation of the 30 lead-lag conditioning click pairs, followed by a test-click pair. When the 30 conditioning click pairs and the test click pair had the same orientation, the echo thresholds were large, indicating a strong build-up. When the test click pair pointed in the mirrored orientation, echo thresholds were shorter. This is in agreement with the idea that *adaptation is direction specific*. If, however, the 30th conditioning click pair pointed to the mirrored location of the previous 29 conditioning clicks, thus indicating a new room configuration, the echo threshold for the test-click pair remained mostly unchanged with respect to the case where all conditioning clicks have the same orientation. In short, one click pair in a different orientation does not completely destroy the adaptation built up by the previous 29, indicating that break-down is not immediate (see also Freyman and Keen 2006).

It makes sense that, if the build-up of echo suppression is not immediate, the break-down should not be immediate either. In the Djelani and Blauert (2002) model, the inhibition algorithm builds up over time through regular and repeated triggering. Therefore, it exhibits the behavior of an integrator or moving average that is affected by past information. It will take a certain number of clicks (i.e. triggers) in the new spatial orientation to *flush out* all of the information from the previous room. In an abstraction process, it is also possible that information about a specific space is integrated over time. If the auditory system has accumulated a lot of information from one space, a new orientation of clicks could take time for the auditory system to resolve and to indicate that the listener has moved to a new environment, particularly in the absence of information from other senses such as vision or proprioception.

Flipping the orientation of the triangular test room in Djelani and Blauert (2001) not only mirrored the direction of the reflections, it also changed the level of each ear signal since individual head-related transfer functions were used which contain interaural level and time differences (ILDs and ITDs). Krumbholz and Nobbe (2002) showed that the ILD in a click pair is more potent for causing a break-down than the ITD, a result confirmed by Brown and Stecker (2013).

Keen and Freyman (2009) investigated the break-down with combinations of a “Room A”, that is, a lead click on the left side followed by a lag click on the right side, and a special “Room B”, which was just the lead click from the left side without the lag. A sequence of Room A click trains increased the echo thresholds for a test click in Room A as expected. If these Room A click trains were followed by an increasing number of clicks from Room B, the echo thresholds gradually decreased with the increasing number of clicks from Room B, eventually reaching the same echo threshold as seen with presentations of Room B clicks only. One interesting point is that it took eleven clicks from Room B to completely break down the build-up caused by five clicks from Room A, indicating the possibility of an asymmetry between the break-down and build-up processes.

While clicks are very useful stimuli in such studies, as they provide quantified amounts of information, many of the stimuli that are encountered in everyday life are more continuous in nature. Therefore, in order to claim that these effects can occur in real-world scenarios, it would be beneficial to see evidence that they also arise with non-transient stimuli. Adaptation and break-down effects were demonstrated for

the Clifton effect paradigm with continuous stimuli, including speech and noise. For example, Djelani and Blauert (2001) presented listeners with a lead from 45° left and a lag from 45° right. After an initial period of at least 3 s in this spatial orientation, listeners pressed a button that caused lead and lag to switch sides immediately. Listeners then reported whether they perceived a temporarily enhanced echo, defined as an echo that either clearly diminishes or fuses with the direct sound again over time. For a train of 2-ms noise bursts at a rate of 4/s, the maximum percentage of temporarily enhanced echoes was reported at delays of 10–20 ms for all listeners. Results were very similar when a speech stimulus was used. However, for a continuous-noise stimulus, only three out of six listeners showed similar results, while two reported hardly any temporarily enhanced echoes and one listener was uncertain. Although noise proved to be a more difficult stimulus for some listeners to detect echoes in a break-down scenario, it is nonetheless clear that break-down does not only occur with clicks.

A study by Clifton et al. (1994) shed further light on the underlying processes. This study looked at discrimination suppression as a function of different lead-lag delays. Listeners were presented with a conditioning train of lead-lag click pairs at a given interstimulus interval, and then a test click pair whose interstimulus interval was varied. For all listeners, discrimination performance for the test click pair with the same lead-lag time delay as the conditioning train was worse, compared to the condition where the conditioning train was not played first, thus confirming the build-up. However, discrimination performance followed roughly a V-shaped pattern, that is, with the performance being worst when the lead-lag delay of the test click pair was identical to that used for the clicks in the conditioning train. Yet, performance improved again for both shorter and longer delays in the test click pair, indicating that discrimination suppression is delay specific. Note that the lag location was mostly unchanged in both situations. This is an interesting result in favor of the abstraction process. When the time delay between lead and lag changes, this could only be caused by a reflecting surface moving either towards or away from the listener, which would be a large change in room geometry. The temporal change would indicate to a stored room geometry model that a large room change has taken place, hence discrimination suppression is shortly reduced. However, in an adaptation process, the location (as implied by the interaural cues) of the lag is suppressed. Consequently, one would assume that only a change in the interstimulus interval does not have such a large, measurable effect on lag-discrimination suppression.

4 From Single Reflections to Room Reverberation

Most of the “rooms” discussed thus far consisted only of a direct sound and a single reflection, whereas typical spaces encountered every day produce a much greater number of reflections. While comparatively unexplored, recent work has begun to investigate build-up and break-down in more complex and realistic room models.

Djelani and Blauert (2001) mimicked the Clifton effect paradigm with a room with a triangular floorplan, which was simulated using the image-source method up to second order, yielding a room-impulse response with a direct sound and 12 reflections. The listeners scaled the size of the triangular room adaptively in order to determine a minimum room size at which the reflections were just barely audible. When the conditioning train and the test click both came from the triangular room in the same orientation, the room needed to be 2–3 times larger in order for reflections in the test click to be perceived, compared to situations where the conditioning room had no reflections or was a left-right flipped room. In all conditions, the source was always placed directly in front of the receiver, so that the leading stimulus was always in the same direction and diotic, whereas the reflection pattern switched ears when the room was flipped.

These results are readily explained by both an adaptation or an abstraction mechanism. While the flipped room will have the same timbral qualities and share the same room acoustic parameters, the spatial locations of the reflections change from the left to the right side and vice versa. This will swap reflections to previously unadapted locations, thereby increasing their audibility in terms of the adaptation model. If the auditory system is storing an abstract geometrical representation of the room, it should likewise be able to identify that the room geometry has changed and, consequently, trigger relearning. One other observation about this particular room is that, in the test orientation, most of the reflections came from the left hemifield, and in the flipped orientation from the right hemifield. Lag suppression could thus build up to one side primarily.

The question remains as to whether the directions of all reflections are suppressed, only some, or if perhaps suppression is weighted by relative amplitude or time after the direct sound. Stecker and Hafter (2002) found that in a click-train stimulus with short inter-click intervals (i.e. <5 ms), the first click was the most strongly weighted perceptually. For longer click intervals, clicks were weighted equally, and when there was a gap in the click train, the first click following the gap was weighted more strongly than those around it. As reflections in a room impulse response are not equally spaced in time, the suppression of any one single reflection could be influenced by the gap in time between itself and the one immediately before it, or by deviation from periodicity.

Rather than looking just at the echo threshold, some recent studies have examined potential benefits of the room-acoustic context on the localization of the direct sound. In the dissertation by Sudirga (2014), localization accuracy was measured in “fixed” rooms, where all trials within a block were simulated from different source locations within the same virtual room, and in “mixed” rooms, where the simulated virtual room varied from trial to trial. There was no significant difference in absolute localization error between the mixed-room and fixed-room paradigms, but the variability in responses was somewhat lower in the fixed-room paradigm—namely, by 1.2–2.4°. In other words, when listeners remained in the same (virtual) room from trial to trial, their localization judgments did not necessarily match the true (i.e. physical) sound source location any better, but localization was more consistent from trial to trial.

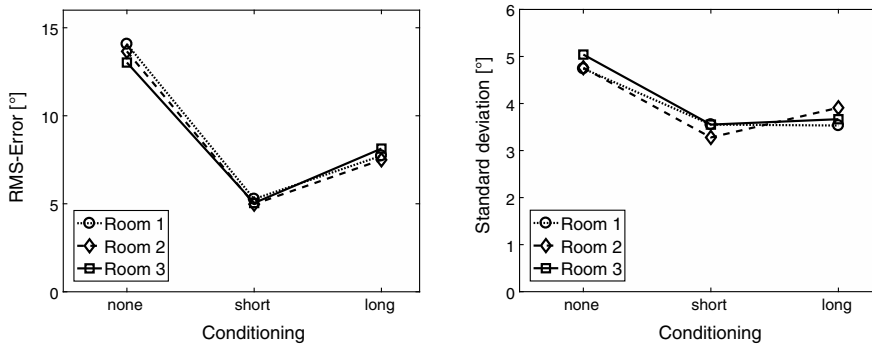


Fig. 4 RMS localization errors, (left), and standard deviations (right), averaged across eight listeners for localization of a click in virtual rooms (1, 2 or 3) when preceded either by silence (conditioning: none), or by a short or a long conditioning sequence of 2 or 14 clicks, respectively. The conditioning clicks were played from random locations left and right of the listeners, while the test clicks were restricted to the non-overlapping frontal region. The stimuli were played in loudspeaker-auralized virtual rooms composed of reflections up to a mirror-image order of 100. The three rooms varied in every trial to reduce across trial build-up. The results show a significant reduction of RMS error and standard deviation when a conditioning sequence preceded the test click. Since the conditioning clicks and the test clicks did not overlap in space, and since test conditions were otherwise identical in all conditions, the observed benefit can be attributed to the context built up by the conditioning clicks. Data replotted from Seeber et al. (2016)

Seeber et al. (2016) tested for direct evidence of room abstraction processes, again through the lens of localization in reverberant environments. Their study examined whether clicks from different source locations within a room improved localization from other source directions in that same room while the listener’s location was kept constant. Such an explicit test for a transfer from one location to another one in the same room, if rendering positive results, could not be explained by an adaptation process because source and reflection positions change with each conditioning click and from the conditioning clicks to the test click. Instead, an improvement due to the room context would support the idea that the auditory system uses interaural cues to infer aspects of the geometry of the room.

In this experiment, a conditioning train of clicks was first played from random locations left and right of the listener in a virtual room. Then, a test click came from a location in the frontal region, and the listener had to indicate its perceived direction. Thus, the conditioning train never contained a click that occurred from the same location or even the region being tested. In this way, the listener has not been exposed to it before completing the localization task—a transfer due to context is required. Three different virtual rooms were used randomly from trial to trial to avoid adaptation from simply hearing an identical room over a block of trials. The test was performed in darkness, in an acoustically treated room.

Figure 4 gives results from eight listeners. Both RMS localization errors and standard deviations of the localization responses were significantly reduced when a conditioning train of clicks (either a short train of 2 clicks or a long train of 15 clicks)

preceded the test click. These results suggest that the conditioning clicks and their reflections create a context in which localization accuracy and precision of subsequent stimuli is improved. Unique to this study is the fact that the conditioning clicks and the test clicks did not come from the same location, and thus a binaural adaptation process cannot explain the improvement. The improvement, therefore, requires a transfer which is in line with some kind of a room abstraction process, particularly as it extends to new locations. Alternately, the context could also be built up by the clicks irrespective of room reverberation, similar to audio-visual contextual effects due to a “visual frame of reference”—compare Radeau and Bertelson (1976). While more research remains to be done in this area, it has become clear that experiments involving more spatiotemporally complex stimuli will be useful to probe the complex processes involved in room adaptation and possibly abstraction in the auditory system.

5 Room Adaptation and Speech Understanding

Understanding speech in reverberant environments is one of the most important tasks performed by the auditory system and the focus of much research activity, particularly for hearing-impaired listeners. Therefore, it is certainly of interest how prolonged exposure to a specific space may improve speech understanding. Research into speech understanding in reverberant spaces has often been studied with respect to the amount of reverberant energy and the shape of the reverberation decay, rather than the specific orientation of direct sound and reflections. Several studies discussed here employ a paradigm with two spoken test words, [sir] and [stir], which differ by the phoneme /t/. In recognition experiments, the amount of reverberation affects the phoneme boundary. In fact, reverberation can make it difficult to differentiate between two similar spoken words because it fills in the gaps that are perceivable in non-reverberant listening conditions.

Watkins (2005) created a continuum of stimuli by interpolating the temporal envelopes of [sir] and [stir] and embedded the respective test words into the context of a sentence. To test the effect of reverberation, the amount of reverberation in the context sentence and for the test words was varied independently. Listeners were played the entire context sentence and then asked whether they heard either [sir] or [stir]. Increasing the reverberation of the test word relative to the surrounding sentence resulted in more [sir] identifications, as there was no opportunity to hear the short glottal stop before the phoneme /t/. When the reverberation of the context sentence was increased relative to the test word, [stir], the number of correct identifications increased. This was interpreted as the reverberation context help uncover the silence before the /t/.

In a subsequent study, Watkins and Makin (2007) used a similar experimental paradigm, but with a noise context rather than a speech context. The same pattern of [sir] versus [stir] identifications as in the previous study could be seen, provided that the noise was broadband and contained temporal modulation in the envelope (i.e.

pauses) that allowed the listener to judge the level of reverberation. Narrow-band noise contexts outside of the relevant frequency bands for [sir]/[stir] identifications did not induce reverberation compensation, even when they had the same temporal envelopes as the broadband noise contexts that did induce compensation. In other words, reverberation compensation appears to be specific with respect to the frequency band.

These results thus help clarify which components of the context are used by the listener to estimate reverberation. A study by Nielsen and Dau (2010) complicates the interpretation since a non-reverberant speech context resulted in more [sir] identifications while other non-speech contexts (including white noise, speech-shaped noise, and amplitude-modulated noise, in addition to silence) with no reverberation to it resulted in more [stir] identifications. This suggests that the non-reverberant speech context itself impedes the detection of the consonant /t/, and that changes in the modulation spectrum of the context are critical for inducing reverberation compensation.

Beeston (2014) has developed a peripheral auditory model to perform the [sir]/[stir]-identification task. The model adjusts efferent suppression in a closed feedback loop based on estimating the amount of reverberation in the pauses of the signal. The model is able to emulate human performance in a [sir]/[stir]-identification task, thereby giving evidence that simple adaptation of peripheral suppression is sufficient for reverberation compensation of speech.

While these studies have investigated the effect of reverberation on a specific phoneme boundary, Zahorik and collaborators have examined more generally how speech understanding in reverberant environments is affected by repeated exposure to the same room acoustics. Brandewie and Zahorik (2010) measured speech reception thresholds (SRTs) in three versions of the same simulated rectangular room, but with different surface materials, to yield strongly different reverberation times. Preceding the target phrase with a sentence carrier and presenting it in the same room showed a significant improvement in SRT by 2.7 dB compared to a no-carrier condition. This is an important result, as it shows that understanding of regular words in noise can be improved by prior exposure to the room. Follow-up experiments with anechoic or monaural stimuli showed a much smaller, non-significant improvement, equivalent to a signal-to-noise improvement of only 0.8 dB. This suggests that the underlying mechanism is binaural and involves a spatial compensation of room reverberation.

Srinivasan and Zahorik (2013) further investigated the time course of this improvement with an open speech corpus and without the sentence carrier. Five different simulated rooms, including an anechoic room, were used to generate the test stimuli, which were presented in either a “blocked” (a block of stimuli all from the same room) or “unblocked” condition (room varied from trial to trial). There was no significant difference in performance between the blocked and unblocked conditions for the anechoic room, while there was a significant improvement in the reverberant rooms when presented in the blocked format. However, no significant improvement was seen over the time course of a group of blocked trials. This suggests that the processes resulting in improved performance operate on fairly short time scales, namely, hundreds of milliseconds to seconds (i.e. within the length of one trial).

It is difficult to tell from these results whether they suggest an adaptation or an abstraction process. Despite their differences in wall absorption, all rooms shared the same geometry, thus leading to the same binaural reflection pattern. An adaptation process should thus suppress the binaural features of reflections similarly in all rooms and show only small recognition differences across conditions, which is in disagreement with present results. On the one hand, an adaptation process could explain the results if one was to postulate that the reflection energy would affect the amount of suppression, which does not easily agree with the above-discussed results on echo thresholds and discrimination suppression, as these are somewhat unaffected by reflection energy. On the other hand, an abstraction process could readily estimate the room geometry in all conditions but would, likewise, have to reset its room-configuration concept upon changes in the absorption characteristics.

6 Self-Motion

In natural settings, people are rarely completely motionless, as, for instance, with their head on a chin-rest. At almost any time, when entering a new space, people will move through a range of different positions within that space. In addition, they are also making small head movements, which have been shown to be important for spatial hearing, particularly with regard to resolving front-back confusions—see, for example, Wightman and Kistler (1999), Thurlow and Runge (1967). Furthermore, listeners often create their own sounds while exploring space, such as with speech or footsteps. When listeners have some sort of input into the room exploration process, this can have an effect on adapting to a new acoustic space versus what can be gleaned from being passively presented with a given stimulus.

Self-motion also has interesting implications for the differentiation of adaptation from abstraction processes. In particular, can the auditory system integrate proprioceptive and vestibular information to update its echo suppression? For instance, if a listener's auditory system has built up suppression to a lead at -45° left and a lag at 45° right, and they turn their head 15° to the left, will they then be adapted to a lead at -30° left and a lag at 60° right (in head-centered, i.e. binaural coordinates)? And if so, is this adaptation or abstraction, assuming that abstraction should be able to deal with all source positions and orientations within a given space? Answers to these specific questions have not yet been given in the literature, but related work on self-motion provides some clues.

Wightman and Kistler (1999) investigated the effect of head movements on resolving front-back confusions and confirmed that they do so. Interestingly, even with a fixed head, front-back confusions were resolved when listeners had control over the movement of the virtual source position (controlled via arrow keys on a computer keyboard), but not when the experimenter controlled source movements—despite both situations being acoustically identical. This suggests that having control over the position of the sound source must work in tandem with the acoustical signals reaching the two ears to show the improvement in spatial hearing. In a similar vein,

Perrett and Noble (1997) showed that head movements also assist with vertical sound-source localization—compare Pastore et al. (2020), this volume.

Echolocation is most famously used by bats to navigate their physical surroundings, but blind and sighted humans can also utilize echolocation after some training. Echolocation requires the listener to produce their own sound stimuli, whereby the direct sound travels directly from the mouth to the ears, and the reflections return from surfaces in the environment. Therefore, it follows that listeners need to have low echo thresholds to echolocate, as the conscious perception of the echoes is what allows echolocation to work. Adapting to a reflection location and having its position suppressed would make it more difficult to echolocate. Thus it might not be surprising that training can affect the sensitivity to ITDs in the lagging sound (Saberi and Perrott 1990).

Wallmeier and Wiegrebe (2014) examined the role of self-motion and head movements in listeners trying to navigate a virtual corridor using only echolocation, with no visual cues. The listeners generated sound stimuli with their own mouths and heard the auralized response from a virtual corridor. They were then asked to orient themselves along the axis of the corridor. Listeners could adjust the orientation of the corridor virtually (like in a video game), adjust the physical orientation of the motorized chair in which they were sitting with head fixed, or adjust the motorized chair with head movements. The ability to move both the chair and one’s head resulted in significantly better performance, indicating the importance of vestibular cues from the motion of the chair as well as from head movements.

This is a scenario where listeners are actually trying to avoid adaptation taking place since they need to be able to hear the echoes distinctly in order to complete the task. An “abstraction” of the room is what the listeners need in order to determine with some certainty where the walls are and in which direction the corridor leads. So, echolocation may be a special case, wherein higher-level cognitive processes attempt an abstraction process to learn the geometry of the space explicitly.

7 Conclusion

This chapter examined the literature for evidence of the processes involved in the build-up and break-down of the precedence effect, starting with pairs of leading and lagging clicks from different spatial locations. This literature overview was then expanded to stimuli with higher numbers of reflections, speech understanding in reverberant environments, and how self-motion can be incorporated into spatial hearing.

The current authors proposed and discussed two possible mechanisms for understanding how the auditory system builds up information to suppress reflections. The first and simpler mechanism has been termed “adaptation” and is affecting binaural and possibly also relevant monaural processes. When determining the location of leading and lagging signals in the binaural auditory system, binaural information from the lagging direction is increasingly suppressed, so that after a matter of sev-

eral seconds, the listener only hears one sound event, rather than two, indicating a rise in the echo threshold. This paradigm has been modeled by Djelani and Blauert (2002) and been shown to predict the “Clifton Effect” (Clifton 1987). This binaural adaptation process might be supported by monaural adaptation processes occurring at the level of hair cells (Hartung and Trahiotis 2001), the cochlear nucleus (Buerck and van Hemmen 2007; Hafter 1996) or through the MOC-feedback loop (Beeston 2014).

The second mechanism, termed “abstraction”, is more complex, but has the potential of being applied to dynamic complex listening scenarios. It is postulated that the auditory system can build up an abstract geometric model of an entire room and use this model to control reflection suppression. This model is predicted to survive, for instance, source and listener movements in the same room, wherein the reflection pattern changes at the listener’s position, but in a way that is consistent with the room geometry. Controlled by proprioceptive information, the use of an abstract room representation avoids having the auditory system re-calibrate every time the reflection pattern shifts, as would have to be assumed for a pure adaptation process. Evidence for this mechanism comes from experiments showing improved localization ability for test positions not part of the exposure sequence but located in the same room (Seeber et al. 2016). A localization improvement of this kind would require a generalization or abstraction of information from the context given by the room rather than by individual reflections.

The work discussed in this chapter provides evidence for both processes in different scenarios. Some results could potentially be explained by either one. Generally speaking, echo suppression and build-up effects can be well explained by an adaptation process while context effects in localization, echolocation and speech understanding in reverberation may suggest an abstraction process. As future work will likely probe more complex and dynamic scenarios, it will help to disentangle the contribution of these different processes or, hopefully, to understand how they might work in tandem to aid listeners when exploring new acoustic environments.

Acknowledgements The present work was supported by the Bernstein Center for Computational Neuroscience Munich, BMBF 01 GQ 1004A and 01 GQ 1004B. SC received funding from the Elite Master Programme in Neuroengineering at the Technical University of Munich. The authors thank Dr. Piotr Majdak and two anonymous reviewers for valuable comments and suggestions.

References

- Beeston, A.V. 2014. Perceptual compensation for reverberation in human listeners and machines. Ph.D. thesis, University of Sheffield, UK.
- Bishop, C., D. Yadav, S. London, and L. Miller. 2014. The effects of preceding lead-alone and lag-alone click trains on the buildup of echo suppression. *The Journal of the Acoustical Society of America* 136 (2): 803–817.
- Blauert, J. 1997. *Spatial Hearing*, 494. Cambridge, USA: MIT Press.
- Brandewie, E., and P. Zahorik. 2010. Prior listening in rooms improves speech intelligibility. *The Journal of the Acoustical Society of America* 128 (1): 291–299.

- Brown, A., and G. Stecker. 2013. The precedence effect: Fusion and lateralization measures for headphone stimuli lateralized by interaural time and level. *The Journal of the Acoustical Society of America* 133 (5): 2883–2898.
- Brown, A., G. Stecker, and D. Tollin. 2015. The precedence effect in sound localization. *Journal of the Association for Research in Otolaryngology* 16: 1–28.
- Buerck, M., and J. van Hemmen. 2007. Modeling the cochlear nucleus: A site for monaural echo suppression? *The Journal of the Acoustical Society of America* 122 (4): 2226–2235.
- Clifton, R. 1987. Breakdown of echo suppression in the precedence effect. *The Journal of the Acoustical Society of America* 82 (5): 1834–1835.
- Clifton, R., and R. Freyman. 1989. Effect of click rate and delay on breakdown of the precedence effect. *Perception & Psychophysics* 46 (2): 139–145.
- Clifton, R., and R. Freyman. 1996. The precedence effect: Beyond echo suppression. In *Binaural and Spatial Hearing in Real and Virtual Environments*, ed. R. Gilkey, and T. Anderson, 233–255. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Clifton, R., R. Freyman, R. Litovsky, and D. McCall. 1994. Listeners' expectations about echoes can raise or lower echo threshold. *The Journal of the Acoustical Society of America* 95 (3): 1525–1533.
- Dizon, R., and H. Colburn. 2006. The influence of spectral, temporal, and interaural stimulus variations on the precedence effect. *The Journal of the Acoustical Society of America* 119 (5): 2947–2964.
- Djelani, T., and J. Blauert. 2001. Investigations into the build-up and breakdown of the precedence effect. *Acta Acustica united with Acustica* 87: 253–261.
- Djelani, T., and J. Blauert. 2002. Modelling the direction-specific build-up of the precedence effect. In *3rd European Congress on Acoustics—Forum Acusticum*, Sevilla, Spain.
- Freyman, R., and R. Keen. 2006. Constructing and disrupting listeners' models of auditory space. *The Journal of the Acoustical Society of America* 120 (6): 3957–3965.
- Freyman, R., R. Clifton, and R. Litovsky. 1991. Dynamic processes in the precedence effect. *The Journal of the Acoustical Society of America* 90 (2): 874–884.
- Hafta, E. 1996. Binaural adaptation and the effectiveness of a stimulus beyond its onset. In *Binaural and Spatial Hearing in Real and Virtual Environments*, ed. R. Gilkey, and T. Anderson, 211–232. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Hafta, E., and T. Buell. 1990. Restarting the adapted binaural system. *The Journal of the Acoustical Society of America* 88 (2): 806–812.
- Hafta, E., and R. Dye. 1983. Detection of interaural differences of time in trains of high-frequency clicks as a function of interclick interval and number. *The Journal of the Acoustical Society of America* 73 (2): 644–651.
- Hartmann, W. 1983. Localization of sound in rooms. *The Journal of the Acoustical Society of America* 74 (5): 1380–1391.
- Hartung, K., and C. Trahiotis. 2001. Peripheral auditory processing and investigations of the “precedence effect” which utilize successive transient stimuli. *The Journal of the Acoustical Society of America* 110 (3): 1505–1513.
- Houtgast, T., and S. Aoki. 1994. Stimulus-onset dominance in the perception of binaural information. *Hearing Research* 72 (1–2): 29–36.
- Keen, R., and R. Freyman. 2009. Release and re-buildup of listeners' models of auditory space. *The Journal of the Acoustical Society of America* 125 (5): 3243–3252.
- Krumbholz, K., and A. Nobbe. 2002. Buildup and breakdown of echo suppression for stimuli presented over headphones—the effects of interaural time and level differences. *The Journal of the Acoustical Society of America* 112 (2): 654–663.
- Lindemann, W. 1986a. Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front. *The Journal of the Acoustical Society of America* 80 (6): 1623–1630.
- Lindemann, W. 1986b. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *The Journal of the Acoustical Society of America* 80 (6): 1608–1622.

- Litovsky, R., H. Colburn, W. Yost, and S. Gunzman. 1999. The precedence effect. *The Journal of the Acoustical Society of America* 106 (4): 1633–1654.
- McFadden, D. 1973. Precedence effects and auditory cells with long characteristic delays. *The Journal of the Acoustical Society of America* 54: 528–530.
- Menzer, F., and B. Seeber. 2014. Does reverberation perception differ in virtual spaces with unrealistic sound reflections? *Proceedings of Forum Acusticum*, 1–4, Krakow, European Acoustics Association.
- Monaghan, J., and B. Seeber. 2016. A method to enhance the use of interaural time differences for cochlear implants in reverberant environments. *The Journal of the Acoustical Society of America* 140 (2): 1116–1129.
- Nielsen, J., and T. Dau. 2010. Revisiting perceptual compensation for effects of reverberation in speech identification. *The Journal of the Acoustical Society of America* 128 (5): 3088–3094.
- Pastore, M., Y. Zhou, and W.A. Yost. 2020. Cross-modal and cognitive processes in sound localization. In *The Technology of Binaural Understanding*, eds. Blauert, J. and J. Braasch, 315–350. Cham, Switzerland: Springer and ASA Press.
- Perrett, S., and W. Noble. 1997. The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America* 102 (4): 2325–2332.
- Radeau, M., and P. Bertelson. 1976. The effect of a textured visual field on modality dominance in a ventriloquism situation. *Perception & Psychophysics* 20 (4): 227–235.
- Saberi, K., and D. Perrott. 1990. Lateralization thresholds obtained under conditions in which the precedence effect is assumed to operate. *The Journal of the Acoustical Society of America* 87 (4): 1732–1737.
- Seeber, B. 2011. The contribution of intrinsic amplitude modulation to the precedence effect at high frequencies. In *Fortschritte der Akustik – DAGA’11*, ed. J. Becker-Schweitzer and G. Notbohm, Dt. Ges. f. Akustik e.V. (DEGA), 833–834, Berlin.
- Seeber, B., and E. Hafter. 2011. Failure of the precedence effect with a noise-band vocoder. *The Journal of the Acoustical Society of America* 129 (3): 1509–1521.
- Seeber, B., M. Mueller, and F. Menzer. 2016. Does learning a room’s reflections aid spatial hearing? In *Proceedings of the 22nd International Congress on Acoustics*, ed. F. Miyara, E. Accolti, V. Pasch, and N. Vechiatti, Buenos Aires, Argentina, ICA2016–775, 1–8.
- Srinivasan, N., and P. Zahorik. 2013. Prior listening exposure to a reverberant room improves open-set intelligibility of high-variability sentences. *The Journal of the Acoustical Society of America* 133 (1): EL33–39.
- Stecker, G., and E. Hafter. 2002. Temporal weighting in sound localization. *The Journal of the Acoustical Society of America* 112 (3): 1046–1057.
- Sudirga, R.E. 2014. Effect of reverberation context on spatial hearing performance of normal hearing listeners. Ph.D. thesis, University of Western Ontario, Canada.
- Thurlow, W., and P. Runge. 1967. Effect of induced head movements on localization of direction of sounds. *The Journal of the Acoustical Society of America* 42 (2): 480–488.
- Wallmeier, L., and L. Wiegrebe. 2014. Self-motion facilitates echo-acoustic orientation in humans. *Royal Society Open Science* 1: 140185.
- Watkins, A. 2005. Perceptual compensation for effects of echo and of reverberation on speech identification. *Acta Acustica united with Acustica* 91: 892–901.
- Watkins, A., and S.J. Makin. 2007. Perceptual compensation for reverberation in speech identification: Effects of single-band, multiple-band and wideband noise contexts. *Acta Acustica united with Acustica* 93 (3): 403–410.
- Wightman, F., and D. Kistler. 1999. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America* 105 (5): 2841–2853.
- Yang, X., and D. Grantham. 1997. Cross-spectral and temporal factors in the precedence effect: Discrimination suppression of the lag sound in free-field. *The Journal of the Acoustical Society of America* 102 (5): 2973–2983.

Room Effect on Musicians' Performance



Malte Kob, Sebastià V. Amengual Garí and Zora Schärer Kalkandjiev

Abstract This chapter reviews the basics of music and room-acoustics perception, an overview of auralization methods for the investigation of music performance and a series of studies related to the impact of room acoustics on listeners and musicians. The acoustics of the performance environment play a major role for musicians, both during rehearsals and concerts. However, systematic investigations of music performance are challenging due to the variety of conditions that determine the artists' performance. Set-ups that allow controlled studies with variable but well-defined acoustic conditions have been developed over the last decades with increasing naturalness and applicability. Current auralization methods allow the reproduction of measured or synthesized room acoustics in real-time, thus enabling the perceptual assessment of room acoustics in laboratory conditions, isolating acoustics from other potential impacting factors. Common methodologies, as well as advantages and limitations of such virtual environments for the study of music and room acoustics perception are discussed in the first section. The virtual environments enable studies that help to explain why and how room acoustics can affect the listener subjective impact of a musical performance and to what extent listeners can be classified depending on their individual taste. Recent studies have shown that musicians systematically adjust their musical performance and adapt to the room acoustical conditions. The most important findings from these studies are presented in the second section. Methods and results from recent investigations of the impact of room acoustics on music performance are discussed in the third section of this chapter.

M. Kob (✉) · S. V. Amengual Garí
Erich-Thienhaus-Institut, Detmold University of Music, Detmold, Germany
e-mail: kob@hfm-detmold.de

S. V. Amengual Garí
e-mail: svamengualgari@gmail.com

Z. Schärer Kalkandjiev
Audio Communication Group, Technical University of Berlin, Berlin, Germany
e-mail: zora.schaerer@tu-berlin.de

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_9

1 Introduction

1.1 Motivation

For the performance of music, the room acoustical properties of the space surrounding the musicians and the audience play a vital role regarding the perception of the sound. The direct sound from the instruments is reflected by the surrounding walls, reaching the performers and listeners with unique reflection patterns and sound characteristics. Musicians depend on the auditory feedback of their performance for expressive fine-tuning (Repp 1999), as well as an internal representation of the sound they intend to convey to the listener (Gabrielsson 1999). It can thus be expected that they adapt their way of playing to the surrounding room acoustics, either the acoustics on stage or the acoustics they assume in the auditorium. This is supported by some of the well-known music treatises of the 18th and 19th centuries (Quantz 1752; Spohr 1833; Czerny 1839) as well as more recent works (Galamian 1962; Borciani 1973) that suggest the use of certain playing techniques related to specific room acoustical conditions. However, in practice, the concepts of how to deal with room acoustics might be different from the theory; adjustments might take place unconsciously, or they might even be entirely rejected (Flesch 1928; Blum 1987).

Beyond seeking empirical evidence of performance adjustments to room acoustics, there are more detailed questions to consider: Are there certain aspects of music performance, such as tempo or dynamic strength, that are adjusted more than others? Which are the room acoustical parameters that influence specific characteristics of the playing technique? Does the way and extent of adjustment depend on factors such as the piece of music that is played, the instrument that is used, or the number of musicians involved?

These questions are interesting from the point of view of music cognition but also of room acoustics. There is ongoing research in stage acoustics concerned with the question of which aspects of the room acoustical environments are relevant for musicians and which physical measures correlate with those venues. Several studies in this field rely on an indirect evaluation of concert halls by using questionnaires distributed to musicians (e.g., Gade 1989b; Sanders 2003; Chiang et al. 2003; Astolfi et al. 2007; Dammerud 2009; Jeon et al. 2015; Panton et al. 2017). However, Gade (2010) pointed out the difficulty musicians have in differentiating between subjective attributes when rating concert halls. This is due to the fact that performers are involved in the production and perception of the music rather than in the analytical evaluation of room acoustics. Furthermore, the vocabulary used to describe room acoustics varies greatly among musicians (Schärer Kalkandjiev 2015, p. 151 ff.). Thus, investigating the direct reaction of performers to room acoustical conditions by observing their changes in playing technique is a promising approach that contributes to stage acoustics research.

1.2 Challenges

Empirical investigations into the influence of room acoustics on music performance face several challenges, outlined in the following.

Music performance involves complex cognitive processes in musicians. Preceding the performance itself, musicians form a concept of how to play the music (Sloboda 1982; Gabrielsson 1999), which is like an overall guide for the sounding enactment. Forming a performance concept involves the acquisition of a mental representation of the music—structural features, emotions, associations, body movements, sound patterns—as well as the practice aimed at attaining the required technical proficiency (Gabrielsson 1999, p. 502 f.). During the actual performance, the sensory (kinaesthetic, tactile, visual, auditory) feedback from the musicians' body movements, the instrument, and the sound guide their playing in order to produce the intended sound. Thereby, both mental and embodied cognitive processing help to activate and control the motor action of playing the instrument. Leman (2008, p. 51) states: “*While involved with music, the human body interacts with physical energy and the human mind deals with interpretations that are built on top of that corporeal interaction.*” Besides this direct interplay of musicians with their instruments, there are several external factors that may influence a music performance. These include the interaction with other performers and the audience, the physical and emotional state of the players, as well as environmental factors such as room size, lighting, room acoustics, and climatic conditions (see Schärer Kalkandjiev 2015, p. 10 ff.). The complexity of the cognitive load, as well as the presence of manifold influencing factors, make it difficult to study the isolated effect of room acoustics on music performance.

Two different approaches exist that can be adopted to investigate the influence of room acoustical surroundings on musicians: (i) conducting field studies in real concert halls or (ii) running laboratory experiments with simulated environments. Both have specific advantages and disadvantages, as pointed out by Gade (2010). Experiments in real halls have a very high degree of external validity, but the experimental variables are difficult to control, and usually, the variation of room acoustical conditions is not large enough. Both problems can be more easily handled in laboratory experiments, but here the realism of the virtual rooms is the major challenge—see Sect. 2.

Empirically studying the performance of music usually involves physical measurements, and in this context, its realization is an essential challenge. This includes defining the most relevant aspects of music performance and selecting those audio features that are most suitable to describe them. Seashore (1938) fundamentally contributed to this issue by defining frequency, amplitude/intensity, duration, and form as the four main physical characteristics of a sound wave with pitch, loudness, time and timbre as their corresponding musical qualities. Many of the studies on music performance that succeeded Seashore's concentrated on specific aspects of these qualities, and a large share of them dealt with piano performances (e.g., see Goebel et al. 2008 for a review). A major impact in this respect can undoubtedly be ascribed to the introduction of the MIDI standard, which immensely facilitated the measure-

ment of piano music in a digital format. The Music Instruments Digital Interface (MIDI) is a standard for a digital protocol and interface among musical instruments, computers and other devices, maintained by the MIDI Manufacturer's Association (1996). Via MIDI, signals are transferred using a serial stream of data with information about note events such as note index/pitch, note on/off, velocity/level and control data. An advantage of MIDI data for music analysis is the immediate availability of digital data describing basic musical events. A drawback of the MIDI standard from 1983 is the condensation of musical expression to a rather small and roughly discretised set of parameters that only represent a part of the original music performance. Enhancements are implemented in the MIDI 2.0 standard from 2019 (midi.org 2020). The small number of studies concerned with other instruments, especially strings, are often confronted with difficulties such as tone, on- and offset detection—see McAdams et al. (2004).

Furthermore, when using either MIDI or audio measurements, careful consideration must be given to the selection of perceptually meaningful physical parameters. One example of this is that the perception of musical dynamics is not only characterized by intensity but also depends on the musical context and timbre (Nakamura 1987). Thus, existing simple loudness measures must be viewed critically when used to describe the dynamic strength of musical pieces—see Sect. 3.2. A further example is that the perception of tempo in music is related to the microstructure of timing in performances (Repp 1994). However, the tempo of performances is often described by merely measuring the duration of whole musical phrases. Due to the lack of a standard operationalization method for music performances, the diversity of tools and methods used in existing studies is huge—see Sect. 3.2—making the comparability among investigations difficult.

It is in the nature of music that there are differences between the performances of the same piece by several players or singers (Sloboda 2000; Gingras 2014; Devaney 2016). The use of certain aspects of performance such as articulation appears to be piece-specific, while other interpretive choices are rather performer-specific (Gingras et al. 2013). Furthermore, there seem to be unintentional and even inaudible, but systematic timing deviations among pianists denoted as “pianistic fingerprint” by Van Vugt et al. (2013). These points raise the questions whether at least certain performative adjustments to room acoustics are performer-specific, as suggested by Schärer Kalkandjiev and Weinzierl (2015), and whether the individual differences in performances are even larger than the differences evoked by the influence of room acoustical surroundings.

1.3 Stage-Acoustics Research

One main concern of room acoustical research is finding physical measures that are suitable to describe the room acoustical perception of audience and musicians. In this context, defining the subjective attributes that are relevant for listeners is a substantial issue that is discussed in the chapter *The Language of Rooms: From*

Perception to Cognition (Weinzierl et al. 2020), this volume. Regarding physical measures, there is a well-established set of parameters that is commonly used to evaluate the auditoria of concert halls (Beranek 2004; Kuttruff 2009; ISO 3382-1 2009) while current research involves new measurement techniques using spherical arrays for sources and receivers and auditory models for the extraction of features (Weinzierl and Vorländer 2015). Concerning the perspective of musicians, however, there are only two standardized physical parameters to describe the acoustics of stage areas, early and late support (ST_{early} and ST_{late}). The following section reviews the progress that has been achieved in the last decades in this field.

Studying different stage configurations on the variable stage of the Gulbenkian Grande Auditorio, Barron (1978) found that all performers of a small ensemble preferred an overhead reflector. In addition, wind players were in favor of close reflecting surfaces around them while strings preferred an open stage. Several authors have studied the benefit of early reflections at certain time intervals, which were judged positively if they arrived

- Between 17 and 35 ms for ensemble musicians (Marshall et al. 1978)
- Before 40 ms for singers (Marshall and Meyer 1985)
- Between 20 and 75 ms for ensemble musicians (Gade 1989a)

Regarding the strength of early reflections, there seems to be a certain limit above which they are disliked by both soloists and ensemble players (Chiang et al. 2003; Ueno and Tachibana 2003; Ueno et al. 2005). Furthermore, the reflection of high frequencies is essential for both instrumentalists and vocalists (Marshall et al. 1978; Marshall and Meyer 1985), and most musicians are in favor of reverberation (Marshall and Meyer 1985; Gade 1989a; Ueno and Tachibana 2003; Ueno et al. 2005).

Regarding specific parameters to characterize the acoustic conditions on stage, Naylor (1988) suggested the measurement of a modulation transfer function with source and receiver on stage, intended to quantify the amount of information conveyed among musicians. After determining the most important subjective properties of room acoustics in an interview study with musicians (Gade 1986), Gade (1989a, b) conducted laboratory and field experiments with ensembles and orchestras that yielded two acoustical stage measures. He found a positive correlation of these measures with both the concepts *hearing oneself* and *hearing others*. After a slight revision (Gade 1992), they are listed in ISO 3382-1 (2009) as ST_{early} —predicting the subjective impression of *ensemble conditions*,

$$ST_{early} = 10 \log_{10} \left(\frac{\int_{20 \text{ ms}}^{100 \text{ ms}} p^2(t) dt}{\int_{0 \text{ ms}}^{10 \text{ ms}} p^2(t) dt} \right), \quad (1)$$

and ST_{late} —predicting the subjective impression of *perceived reverberance*,

$$ST_{\text{late}} = 10 \log_{10} \left(\frac{\int_{\frac{100 \text{ ms}}{10 \text{ ms}}}^{1000 \text{ ms}} p^2(t) dt}{\int_{0 \text{ ms}} p^2(t) dt} \right), \quad (2)$$

with the sound pressure, $p(t)$, measured at a distance of 1 m between source and receiver. Both transducers are mounted at a height of 1 m or 1.5 m (ISO 3382-1 2009).

Low correlation values between perceptual properties of stage acoustics and the support parameters were found in later studies (O'Keefe 1995; Chiang et al. 2003; Berntson and Andersson 2007; van Luxemburg et al. 2009; Dammerud et al. 2010). Therefore, there were suggestions for alternative measures (Chiang et al. 2003; van den Braak and van Luxemburg 2008; Brunskog et al. 2009; Dammerud 2009; Wenmaekers et al. 2012) of which two promising approaches shall be mentioned here: The parameters G_e and G_1 are presumably less prone to effects of the source directivity, the distance between source and receiver as well as the floor reflection. They can be calculated from the in-situ measurements of sound strength G and clarity C_{80} (Dammerud 2009).

$$G_e = 10 \log_{10} \left(\frac{\int_{0 \text{ ms}}^{80 \text{ ms}} p^2(t) dt}{\int_{0 \text{ ms}}^{\infty} p_{10 \text{ m}}^2(t) dt} \right) = 10 \log_{10} \left(\frac{10^{C_{80}/10} \cdot 10^{G/10}}{1 + 10^{C_{80}/10}} \right), \quad (3)$$

$$G_1 = 10 \log_{10} \left(\frac{\int_{\frac{80 \text{ ms}}{\infty}}^{\infty} p^2(t) dt}{\int_{0 \text{ ms}}^{\infty} p_{10 \text{ m}}^2(t) dt} \right) = 10 \log_{10} \left(\frac{10^{G/10}}{1 + 10^{C_{80}/10}} \right), \quad (4)$$

with the sound pressure, $p(t)$, of the in-situ measured impulse response and the sound pressure of an impulse response measured in the free field with the same sound source at a distance of 10 m to the receiver, $p_{10 \text{ m}}(t)$.

Wenmaekers et al. (2012) introduced an extension of ST_{early} and ST_{late} so that these parameters can be measured with source-receiver distances larger than 1 m. A greater distance allows positioning the source and receiver at the location corresponding to different instruments, and thus accounting for the benefit of early and late reflections when it comes to hearing the sound of neighbored instruments on stage:

$$ST_{\text{early},d} = 10 \log_{10} \left(\frac{\int_{10 \text{ ms}}^{103 \text{ ms} - \tau} p_d^2(t) dt}{\int_{0 \text{ ms}}^{10 \text{ ms}} p_{1 \text{ m}}^2(t) dt} \right), \quad (5)$$

$$ST_{\text{late},d} = 10 \log_{10} \left(\frac{\int_{10 \text{ ms}}^{\infty} p_d^2(t) dt}{\int_{0 \text{ ms}}^{103 \text{ ms} - \tau} p_{1 \text{ m}}^2(t) dt} \right), \quad (6)$$

sound pressure measured with a distance of 1 m between source and receiver, $p(t)_{1 \text{ m}}$ the sound pressure measured with arbitrary distance d between source and receiver, $p(t)_d$ and the delay between source and receiver, τ , in milliseconds.

As a perspective for future work, several authors have advocated the use of directional sources and receivers for the measurement of stage acoustical parameters (Meyer and Biassoni de Serra 1980; O'Keefe 1995; Dammerud 2009; Wenmaekers et al. 2017). In addition to monaural parameters, the use of microphone arrays allows for the exploration of spatial properties of sound fields on stage (Guthrie et al. 2013; Panton et al. 2016). Because of the presumably large number of variables in stage acoustical investigations, it was argued that especially in field experiments the number of studied concert halls needs to be quite large to avoid the confounding of variables (Gade 2010). Dammerud particularly notes occupied stage conditions (chairs, people) as an element impacting realism (Dammerud et al. 2011). In order to achieve more comparable results, there have been efforts to establish an absolute uniformity regarding the measured acoustical and architectural parameters and the questionnaires used for collecting subjective data (Gade 2013).

2 Auralization Applied to Music Performance

Auralization is the process of rendering and delivering audible soundfields to listeners, recreating the acoustic impression of a real or simulated environment. Either indoor (room acoustics) or outdoor (soundscapes) environments can be auralized, and a vast number of measurement, simulation, and reproduction techniques are available to this endeavour (Kleiner et al. 1993). This section provides an overview of common techniques used to auralize room acoustics in real-time for their application to the study of music performance.

The basic steps to auralize room acoustics consist of convolving anechoic live or recorded sound with a set of filters generated from simulated or measured spatial room impulse responses (SRIR) and reproducing the resulting signals using either a loudspeaker set-up or a pair of headphones. A block diagram of the process is depicted in Fig. 1. The star symbol denotes the convolution of anechoic signal and spatial impulse response.

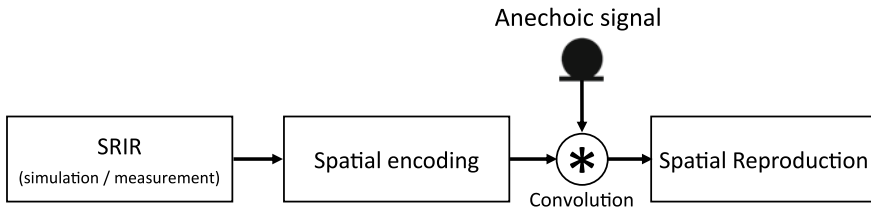


Fig. 1 Basic steps to auralize room acoustics

2.1 Transfer Path

Room acoustics contributes to the perception of the musical performance of listeners in the audience as well as of the players. Whereas the acoustic path from an instrument to listeners is mostly unidirectional (with the exception of wanted or unwanted reactions of the audience), musicians generally react to and even depend on the acoustic perception of their own performance. In Fig. 2, four different sound transmission paths are indicated for the case of a speaker or singer and a listener in the audience. The direct sound between singer and listener corresponds to the propagation under free field conditions, i.e., without any influence of bounding walls or obstacles. This path is determined by the radiation characteristics of the musical instrument and the hearing of the listener. The arrival time of early reflections follows within 5–100 ms after the direct sound. These components essentially contribute to the perception of room properties, support the location of the musician but can also amplify the direct sound (Barron 1974). The sound field components that are perceived after these early reflections are attributed to the diffuse field in the room. They do not support the localization of the musicians, nor do they support the direct sound. Instead, these components add to the listeners' feelings of envelopment and raise the overall sound level of the performance (Kuttruff 2009; Griesinger 1997). The fourth sound path is the self-perception of musicians. This path could be further divided into an extraaural path via the air and room response and the sound path within the musician's (bone conduction) and instrument's bodies. Whereas the internal sound path can be very fast due to the high speed of sound within the human body or the body of the instrument, the path through the air includes all three types of sound paths between music instrument and listener and therefore extends over a large time scale. Another difference between these paths is the wave type; while airborne sound always travels as a longitudinal wave, bone conduction and sound propagation in instruments can also use transversal and bending waves, resulting in more complex phenomena such as dispersion and early vibratory or tactile perception of sound (Sarvazyan et al. 2013).

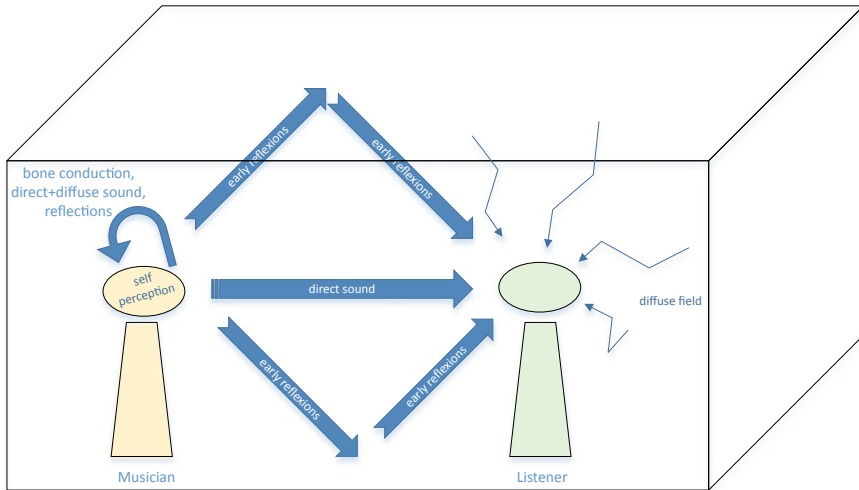


Fig. 2 Transfer paths from musician to listener, including self-perception and diffuse field due to late reflections

2.2 Generation of Spatial Room-Impulse Responses

A Spatial Room Impulse Response (SRIR) can be regarded as the *acoustical fingerprint* of a room, describing its time-energy and spatiotemporal acoustical behavior. Thus, in order to create plausible and/or realistic auralizations of rooms, it is crucial to obtain an SRIR of the target room, describing the acoustical transfer path from the source to the receiver position. In the specific case of musicians on stage, the source position is located at the instrument position, and the receiver corresponds to the musician's ears.

Two major approaches are available in order to obtain an SRIR: measurements and simulations. Measurements are mainly used to resynthesize the acoustic conditions of real rooms in laboratory conditions. When geometry and sound absorption properties of a room are known, simulations can be used as well to replicate the acoustics of real rooms. In addition, simulations allow a high degree of flexibility, as the absorption properties of the materials and the geometry of the room can be modified.

Standard acoustical measurements in performance rooms, as described in the ISO standard (ISO 3382-1 2009), are typically conducted by using an omnidirectional source and microphone. This results in the estimation of monaural room parameters, such as Reverberation Time (RT_{60}) or Clarity (C_{50} , C_{80}), among others. However, these measurements are not suitable for auralization, since a monaural room impulse response (captured by a single microphone) provides only time-energy information, and spatiotemporal information is needed in order to create more realistic auralized versions of measured rooms. Then, arrays composed of multiple microphones are needed to capture SRIRs. The topology and dimensions of a microphone array

are determined by the analysis technique used to extract spatiotemporal information from SRIRs. Common techniques to encode measured sound-fields are Higher Order Ambisonics (Gerzon 1973; Malham and Myatt 1995; Daniel 2003), Directional Audio Coding (Pulkki 2006), Spatial Decomposition Method (SDM) (Tervo et al. 2013), Wave-Field Analysis (Berkhout et al. 1997), or binaural technique (Blauert 2013), among others. A straightforward approach to generating a Binaural Room Impulse Response (BRIR) consists of capturing a stereo room impulse response using an artificial head with microphones placed inside the ear canal.

Room acoustic simulation techniques can be classified into two main groups: geometrical acoustics (GA) techniques and wave-based methods. Geometrical techniques assume wave propagation can be represented by rays, which is a valid approximation for high and mid frequencies. For low frequencies, wave-based methods are more appropriate, as they solve the discretized wave equation numerically and are able to replicate wave phenomena such as room modes and diffraction. Using wave-based methods is computationally very expensive. Therefore combinations of GA and wave-based techniques can be used to simulate the entire audible range. Common GA techniques are Ray Tracing (Ondet and Barbry 1989), Mirror-Image Source (Allen and Berkley 1979; Borish 1984), or Beam Tracing. Savioja and Svensson (2015) presented an extensive review on state of the art regarding GA. Finite-Difference Time Domain (FDTD; Botteldooren 1995) and Finite-Elements Method (FEM; Pietrzyk and Kleiner 1997) are standard techniques to implement wave-based simulations. Apart from the expensive computational costs, results from simulations do not typically match the realism of auralizations based upon measured impulse responses, much due to the challenges of characterizing material acoustic properties (Brinkmann et al. 2017, Brinkmann et al. 2019).

2.3 *Sound-Field Reproduction*

To auralize a measured or simulated SRIR, it is necessary to reproduce the spatialized soundfield, by either loudspeaker set-ups or headphones. The main operation consists of generating appropriate auralization filters and convolving the SRIR with the live or recorded sound generated by a musician.

A BRIR can be directly convolved with recordings and live sound, and then reproduced using headphones. The use of individualized head-related transfer functions (HRTF) provides benefits in localization, especially in the median plane (Wenzel et al. 1993; Møller et al. 1996). However, the amount of individualization required to achieve plausible auralization of room acoustics is a topic under investigation (Begault et al. 2001). In addition, headphone equalization should be considered (Schärer and Lindau 2009; Brinkmann and Lindau 2012). Moreover, the case of dynamic binaural synthesis requires tracking the head movements of the listener and updating the BRIR accordingly.

Reproducing 3D sound-fields using loudspeakers is usually computationally more demanding, and multichannel set-ups can be fairly expensive. However, listeners are

freed from wearing headphones, and it is not necessary to track their movements to reproduce the appropriate soundfield. Common techniques to reproduce 3D soundfields are Vector Base Amplitude Panning (VBAP) (Pulkki 1997), Ambisonics (Gerzon 1973), Wave-Field Synthesis (Berkhout et al. 1993)—although in practice it is typically implemented in a 2D configuration, or Nearest Loudspeaker Synthesis (Tervo et al. 2015), among others. A hybrid technique is Dynamic Cross-Talk Cancellation (CTC), where binaural signals are reproduced through loudspeakers (Lentz 2006).

2.4 *Real-Time Auralization*

The core operation of real-time auralization consists of real-time convolution between the live sound generated by a performing musician and the auralization filters created from a measured or simulated SRIR appropriately treated using a spatial reproduction method. Partitioned convolution schemes are efficient techniques for real-time convolution (Gardner 1995).

The total system latency is defined from the moment of generating a sound using a musical instrument until the corresponding room reflections are reproduced. Minimizing the latency is crucial to allow real-time interaction of musicians with the system. In some cases, if the latency is too large, some reflections (e.g., first-order floor reflection, arriving at approximately 6 or 9 ms for seated or standing musicians, respectively) may not be reproduced. If binaural reproduction is used, a hard floor surface can be included to create a floor reflection physically.

The level difference between the direct sound of the instrument and the artificially auralized reflections should be calibrated as well. However, given that instrument directivity is a complex issue and in simulation or measurements source radiation is commonly simplified, there is not a straightforward approach to perform this calibration. Amengual Garí (2017) and Schärer Kalkandjiev (2015) describe approaches for the calibration of loudspeaker and binaural auralizations, respectively.

2.5 *Available Virtual-Acoustic Environments*

This section presents an overview of virtual acoustic environments that have been used over the last decades to study live music performance. Each of the reviewed environment uses different techniques to generate and reproduce SRIR, showing that multiple approaches can be combined. However, this results in a lack of standardization in evaluating the performance of auralization techniques, thus difficulting the comparison and experimental repeatability of research results in this field.

The first virtual acoustic environment used to conduct research related to a live music performance was implemented by Marshall et al. (1978). However, in this case, the auralization consisted only of a limited number of reproduced early reflections,

without late reverberation. Later on, Gade (1989a) implemented an environment which reproduced early reflections and late reverberation. The live sound of the musician was captured, and the early reflections were simply delayed versions of the live sound. To reproduce late reverberation, the captured sound was reproduced in a reverberant chamber to be then captured again and reproduced to the musician. These environments represent simplified and limited approaches, and with the implementation of digital signal processing (DSP) techniques the capabilities and plausibility of auralization environments improved substantially.

Ueno et al. (2001) implemented an auralization system based on the measurement and loudspeaker reproduction of 6-channel measured SRIR. The measurement set-up consists of an omnidirectional loudspeaker and a directional microphone placed on a stage, obtaining 6 directional impulse responses by rotating the microphone (2 for each orthogonal axis). Then, the sound of a musician performing in an anechoic chamber is captured by a microphone and convolved with the measurement responses. Although only equalization and no spatial treatment of the measured SRIRs are involved, the monaural characteristics of the real and auralized environments showed considerable agreement. However, the spatial properties of the auralized sound-fields were not analyzed. Later, Ueno et al. (2005) extended the same principle to simultaneously auralize soundfields in real-time corresponding to a duo on stage.

The Virtual Singing Studio (Brereton et al. 2012) is an environment specifically designed to study singing voice performance. The system consists of a 16 loudspeaker set-up reproducing the result of convolving measured first-order Ambisonics SRIR with the live sound of singers. The authors also studied the effect of different microphone models and positions.

The Virtual Performance Studio (VPS) (Laird et al. 2011) is an acoustic virtual environment based on simulation of room acoustic models and 12 channel loudspeaker reproduction of first-order Ambisonics SRIR. The authors make particular emphasis on the issues that musicians can encounter when performing in an acoustic virtual environment, such as the proximity effect (increase of low-frequency content) caused by close-miking techniques, the PA-effect (amplification of non-musical sounds and tonal distortion) or the restriction of musicians' movements when performing due to microphone and loudspeaker positioning.

Second-order ambisonics measurement and reproduction of SRIR is used by Guthrie et al. (2013) to deliver real-time auralization of room acoustics to musicians. The auralized soundfields are reproduced using a 3D set-up of 18 loudspeakers and 4 subwoofers. The main application of this environment is the study of spatial parameters influencing stage acoustic preferences.

In order to conduct music performance studies in controlled acoustic conditions, Schärer Kalkandjiev (2015) implemented an auralization system based on binaural reproduction of simulated room acoustics. The generation of SRIRs is produced by a combination of an image-source model and ray tracing. In addition, multiple simulations of each room are implemented using different source radiation models, in order to fit the room excitation properties to those of different instruments. From these simulations, a database of BRIRs corresponding to different head orientations

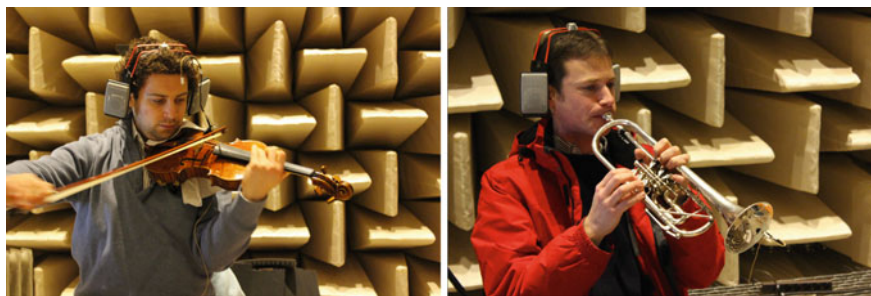


Fig. 3 Musicians performing in a binaural environment using non-occluding headphones. Image extracted from Schäfer Kalkandjiev (2015)



Fig. 4 Trumpet players performing in the D3S environment. Image extracted from Amengual Garí (2017)

is generated and used later for real-time convolution with the live sound of musicians in an anechoic chamber. The reproduction is based on dynamic binaural resynthesis using head tracking and thus allowing musicians to move during the performance freely. In order to not block the path of the direct sound from the instrument to the musicians' ears, extra-aural (non-occluding) headphones are used (Fig. 3). The problem of needing multiple measurement or simulation sets to appropriately account for source directivity can be solved by using a microphone array to capture the live signals. This was demonstrated in practice by Arend et al. (2019), using a microphone array of 32 microphones in a similar binaural virtual environment.

The Detmold Surround Sound Sphere (D3S) is a loudspeaker based environment for real-time auralization of measured SRIR (Amengual Garí 2017). The process consists of stage measurements using a directional sound source and a 3D microphone array, positioned at the instrument and musician's head location in order to approximate the room excitation to a real performance situation. The captured SRIR are analyzed using SDM, and auralization filters are synthesized using VBAP. The re-synthesized SRIR are convolved in real-time with live sound captured by a close directional microphone, and the result is played back using a 3D loudspeaker set-up composed of 13 loudspeakers. Physical measurements have validated the auralization quality, showing that the time-energy properties of the auralized sound-field agree within ± 3 dB with the original room in the range 200 Hz–4 kHz. Also, the directional properties of the auralized and the original sound-field present considerable agreement.

The total round-trip delay from the live input to the musician's ears is 8.7 ms, allowing the reproduction of the first order floor reflection. A proof of concept has been tested with trumpet players, conducting systematic tests on music performance and stage acoustic preferences (Fig. 4).

3 Impact of Room Acoustics on Music Performance

The previous sections have provided some fundamentals of methods for implementation of acoustically controlled environments for music performances. This section shall elucidate approaches for the evaluation of musical performance and some results obtained from their application.

3.1 Music-Performance Analysis (MPA)

Music Performance Analysis is a subfield of Music Information Retrieval (MIR) (Downie 2003), which aims at extracting information from audio or symbolic (MIDI) signals, in order to characterize musical aspects of a musical performance.

Similarly to a sound wave, musical aspects can be categorized in terms of amplitude, time, pitch, and timbre. However, each of these groups is composed of multiple individual aspects e.g., time is related to rhythm, speed, and articulation. Besides, the modulation of each musical aspect leads to a musical result, which is interpreted in a highly distinctive way by both musicians and listeners. For instance, the characteristics of amplitude modulations in a performance (modulation frequency, minimum and maximum level, micro and macro modulations) are part of musical dynamics and, from a musical perspective, the definition and quantification of dynamics acquire a partially subjective significance.

Conventional approaches to MPA consist of extracting low-level audio or MIDI features, which provide objective information about the analyzed performance. An example of the simplest low-level descriptors for the previously mentioned categories could be root-mean-square sound level (amplitude), the total duration of a performance (time), median fundamental frequency (pitch) or spectral centroid (timbre). For the case of MIDI recordings, low-level features can be easily extracted from the encoded data, given that the duration, pitch, and velocity parameters of each recorded note are available. The extraction and combination of multiple features allow the construction of more complex descriptors that aim at characterizing the human perception of music in terms of musical and emotional aspects (Friberg et al. 2011).

Several tools allowing the extraction of performance features from audio and MIDI recordings are currently publicly available. Some of these tools are the MIR and MIDI toolboxes for Matlab (Lartillot and Toiviainen 2007; Eerola and Toiviainen 2004), the jMIR package (Mckay and Fujinaga 2009), the stand-alone software Sonic

Visualiser with an extensive collection of plug-ins (Cannam et al. 2010), or the CUE EXtraction (CUEX) algorithms (Friberg et al. 2007). By no means this list is meant to be complete, but rather a starting point for the reader towards tools previously used for research. In many cases, it is common to combine multiple tools into more complex models or opt for a full custom implementation of several feature extraction algorithms (Lerch 2008; Amengual Garí 2017).

3.2 *Performance Adjustments of Musicians*

Several musicians and music scholars have given specific recommendations regarding the playing techniques to be used in certain room acoustical environments. They can be summarised as follows:

- Avoid fast tempi in large, reverberant halls because tones are otherwise blurred (Quantz 1752; Galamian 1962)
- Do not force the tone of the instrument (especially strings) in both large and dry halls (Spohr 1833; Galamian 1962; Borciani 1973)
- Instead of forcing the tone (see above), increase the amount of sound by dividing bow strokes in large, dry halls (strings; Borciani 1973; Galamian 1962)
- Use a wide vibrato in large, dry halls (strings; Borciani 1973)
- Prolong notes in dry halls (Blum 1987)
- Articulate clearly in reverberant halls (Blum 1987)

In the search for evidence showing the use of these or other adjustment strategies by musicians, there have been a growing amount of empirical investigations on this topic. Structured by the two basic approaches of field and laboratory studies, the following section reviews the research and main findings in this field.

Field Studies

Winckel (1962) carried out an early field experiment with the Cleveland Orchestra directed by George Szell, who performed 8 musical works in 15 concert halls in the USA. To characterize the room acoustics, Winckel recorded the decay of a musical chord in the occupied and unoccupied concert halls (RT between 1 and 2.1s). As for performance properties, Winckel measured the maximum and minimum sound pressure levels (SPL) of several musical phrases in the audience area as well as the duration of each movement using a stopwatch. The SPL measurement revealed a similar upper dynamic limit in most rooms. Assuming that the concert halls did not have the same sound strength and given the fact that the measurement took place in the audience area, a constant maximum SPL implies an adjustment of the dynamic strength by the orchestra. Since the minimum SPL of the orchestra was more variable across rooms, Winckel assumed that it was mostly influenced by the background noise level of each space. The playing tempo of the musicians did not correlate with RT as one might expect concerning the recommendations by music scholars. However, Winckel concluded that the orchestra used the slowest tempi in

those halls with very good hearing conditions and the fastest tempi in rooms that were unsuitable for live music performances. This implies that the playing tempo is linked to the perception of room acoustical quality.

A more recent investigation was conducted with a well-known cellist who was recorded during his performances of the *Six Suites for Violoncello Solo* by J. S. Bach in seven European concert venues (Schärer Kalkandjiev and Weinzierl 2013). These halls were simulated in computer models in order to obtain room acoustical parameters on each stage by using a source with the directivity of a cello. Technical features were extracted from the audio recordings and then used as predictors in perceptually-based regression functions to determine temporal, dynamic and timbral performance properties (see Schärer Kalkandjiev 2015, p. 51 ff. for details). The statistical analysis used to study the effect of four room acoustical parameters on seven performance characteristics yielded an explained variance of more than 50%. Remarkably, such a high share of the performance variance was due to the room acoustical conditions since there are many other influencing factors in real-world concert situations (see Sect. 1.2). An examination of specific interrelations showed that, as in Winckel's study, there was no linear correlation between RT and the tempo of the cellist. Instead, he played significantly slower in rooms with both short and long reverberation times, and this effect was stronger for the fast movements of the cello suites. The dynamic strength of the cello performances was also significantly reduced in rooms with short and long reverberation times (Schärer Kalkandjiev 2015). This is interesting because it partly contradicts the results of other studies that found a negative linear correlation between RT and the dynamic strength of performers (von Békésy 1968; Bolzinger et al. 1994). However, interviews conducted with the cellist shed some light on this finding: Apparently, he had learned to play soft when encountering a lack of acoustical liveliness instead of forcing the sound, thus complying with the recommendations of music scholars (Spohr 1833; Galamian 1962; Borciani 1973). At the same time, he felt the need to hold back with the increasing reverberation time, which is in turn in line with the previous studies. In terms of timbral adjustments, the results showed that the cellist played significantly harder and brighter in rooms with high late support (ST_{late}), a measure for the perceived reverberance on stage (ISO 3382-1 2009). A hard and bright tonal rendition is likely to be related to playing *trenchant* and to using a more defined attack in articulation, both reported as strategies adopted in very diffuse environments by the cellist. It should further be mentioned that the strongest effects in the statistical analysis were found for the influence of the reverberation time and late support on timbral performance attributes (Schärer Kalkandjiev 2015). This firstly emphasizes that it is important to consider timbral attributes when studying the effect of room acoustics on music performance and secondly it indicates that these two room acoustical parameters are relevant for the perception of solo musicians on stage.

Laboratory Studies

In an early laboratory experiment with trained and untrained pianists von Békésy (1968) asked the musicians to play pieces of varying difficulty and familiarity in three rooms. He measured the reverberation time to represent the room acoustics

(3.8, 1, and 0.6 s) and the vibration amplitude of the piano body as a measure for the dynamic strength of the performances. In the case of one pianist, the vibration amplitude negatively correlated with *RT* while the maximum dynamic range was found in the intermediate reverberance condition. In addition, the least pronounced adjustments were found for unprofessional pianists playing difficult and unfamiliar pieces. Most likely in these cases, the musicians were cognitively absorbed by mastering the technical difficulty of the pieces—part of forming a performance plan, as mentioned in Sect. 1.2—so they could not concentrate on adapting to the room acoustical conditions. Further experiments were conducted by Bolzinger et al. (Bolzinger and Risset 1992; Bolzinger et al. 1994) in a room treated with four different configurations of absorbing material, featuring reverberation times ranging from 0.3 to 1.5 s. General trends among pianists were observed: when playing in more reverberant conditions, musicians tended to decrease the amount of sustain pedal and the overall performance level. The extraction of performance data was done using a MIDI interface equipped on a piano.

Pipe-Organ experiments conducted in a concert hall with electronically enhanced acoustics concluded that also organists systematically adjust their performance to suit acoustics (Amengual Garí 2017, Amengual Garí et al. 2015). Three reverberation settings were implemented, increasing the reverberation time of the hall significantly, in an attempt to replicate acoustics more suited to those typically present during organ performances. Five music students were asked to prepare a few musical excerpts and recorded under different conditions using a MIDI interface. Given that organ dynamics are constant and mostly depend on the registration, the analysis focused on temporal aspects of the performance. As a general rule, musicians tended to decrease the playing tempo in increased reverberation settings, but it was observed that the degree of adaptation greatly depended on the nature of the piece. When playing pieces with loud and full registrations, featuring strong chords and prominent breaks, players were more inclined to systematically adapt their performance, decreasing the overall tempo and increasing the length of long breaks. Contrarily, pieces with soft registration and legato articulations were not significantly adapted.

Ueno et al. (2007) made use of the increasing naturalness and plausibility of room acoustical auralization systems by simulating four performance spaces in an anechoic chamber with 6-channel loudspeaker reproduction. Five solo musicians (violin, oboe, flute, vocals) were recorded while performing the same two phrases (excerpts of a fast and a slow piece) in each room. Technical features were extracted from the audio recordings, namely phrase duration, A-weighted SPL, fundamental frequency fluctuations, SPL fluctuations and spectral features, in order to describe the performance characteristics tempo, dynamic strength, vibrato and timbre/articulation. The technical features all varied with the room acoustical conditions, but in most cases, the manner of adjustment depended on the instrument (Kato et al. 2007). However, the players' individuality likely accounted for the different adjustment strategies (see Sect. 1.2), since except in one case there was only one musician for each instrument. Similar to the results of Schärer Kalkandjiev and Weinzierl (2013), a consistent reaction among the musicians was to reduce the tempo—especially when playing the fast piece—in both the most reverberant and the anechoic room. The latter is

supported by interviews conducted with the musicians in which they explained that they tried to prolong notes in acoustically dry environments (Ueno et al. 2010), a reaction that most probably reduces the overall tempo of a performance. Some performers reduced their dynamic strength in both the anechoic and most reverberant room (Kato et al. 2007), as it was reported for the cellist in the field study outlined above (Schärer Kalkandjiev 2015). At the same time, an investigation of the effect of specific room acoustical parameters on the mean dynamic strength of the performances revealed a significant negative correlation with the support parameters ST_{early} and ST_{late} (Ueno et al. 2010). Regarding the timbre and articulation of performances in reverberant rooms, the strategies found by Kato et al. (2015) were suppressing higher harmonics, prolonging pauses between notes and using a more pronounced *staccato*. Along similar lines, playing shorter notes in reverberant rooms was a strategy described by many performers in the interviews (Ueno et al. 2010).

Schärer Kalkandjiev and Weinzierl (2015) carried out an experiment in an anechoic chamber with 12 professional solo musicians of 6 instruments (violin, cello, clarinet, bassoon, trumpet, trombone) playing two pieces (slow and fast) in 14 virtual concert spaces simulated with dynamic binaural synthesis (see Sect. 2.5). The methods for determining room acoustical measures and performance characteristics were the same as in the field study described above (Schärer Kalkandjiev and Weinzierl 2013). The statistical analysis showed that five room acoustical parameters accounted for only 2% of the variance of eight performance attributes averaged over musicians. However, if the individual adjustment strategies of the players were taken into account, the explained variance increased to 13%. Firstly, this result demonstrates how large the impact of musicians' individuality can be (see Sect. 1.2). Secondly, it is remarkable that the equally calculated explained variance in the field study was almost four times higher than in the laboratory study, despite many other influential factors. It seems that the absence of visual information about the concert halls in the laboratory study did not aid the musicians' concentration on adjusting to the room acoustics. Instead, much attention was drawn by the effort to get a mental image of the simulated rooms, as interviews revealed. Furthermore, visual and acoustical properties of rooms usually covary so that they may have a stronger effect as an entity. When conducting laboratory studies, it is thus not only the plausibility of the acoustical simulation that needs to be taken into account (Gade 2010) but more precisely the ability of musicians to engage with the simulation as a concert-like situation, possibly including visual cues.

Turning to the specific interrelations revealed in this laboratory study, a significant negative correlation was found between RT , and the tempo of the slow pieces averaged over musicians. This effect was dominant for fast pieces in previous studies (Kato et al. 2007; Schärer Kalkandjiev and Weinzierl 2013), and interviews conducted with the performers explained this result: Most of the musicians mentioned that they concentrated on a precise articulation instead of slowing down the tempo when playing fast pieces. It can be concluded that there are different strategies on how to react to reverberant room acoustical conditions. The choice of strategy seems to depend on the basic tempo of the piece, but the musical character and the musician's individuality also appear to influence the selection. In accordance with Spohr (1833), Borciani (1973) and Galamian (1962) and just as musicians in other studies

(Kato et al. 2007; Schärer Kalkandjiev and Weinzierl 2013), the players significantly reduced their dynamic strength with decreasing reverberation time. The tonal rendition of the performers was affected most strongly by room acoustical parameters (Schärer Kalkandjiev 2015). At the same time, significant differences between the performative adjustments of some instruments—especially regarding the timbre characteristics—were revealed, so an awareness of the played instrument is necessary when investigating the adjustment of timbre properties. Comparing the five room acoustical predictors with respect to their impact on each individual musician's performance showed that the stage parameters G_e and ST_{late} had the greatest influence.

A further point examined by Schärer Kalkandjiev (2015) was the influence of perceived room acoustical quality on the performance of music. After collecting individual quality ratings of the simulated concert halls from each musician, it was shown that on average the musicians played significantly slower (see Winckel 1962), with more dynamic strength and with increased dynamic and timbral bandwidth in rooms they liked. In the interviews, the performers referred to a reduction of tempo and free use of dynamics under favorable conditions as well as fast playing in rooms they did not like.

Systematic studies with 11 trumpet musicians were conducted by Amengual Garí (2017) and Amengual Garí et al. (2019). In these studies, the D3S virtual acoustic environment was used to reproduce measured acoustics of different rooms at the Detmold University of Music. These studies consisted of preference ratings of stage acoustics and recording sessions, including personal interviews. The recorded performances were then analyzed to extract performance features. Given the high dimensionality of the extracted performance data, a Dual Multiple Factor Analysis (DMFA) was conducted on the dataset, reducing the data dimensionality to 4 main dimensions: Overall level and timbre, dynamic variations, overall tempo, and tempo variations. The results suggested that most of the players reduce the overall level and produce a darker timbre when performing in more reverberant and energetic environments. All the players adjusted their performance to a certain degree, although some musicians were prone to implement greater adjustments. A cluster analysis of the experimental results showed that classifying musicians depending on their performance adjustments is not straightforward, due to the multidimensionality of the performance analysis. However, musicians can be classified with regard to single dimensions, and the adjustments can be partially categorized. Similar results were found by Luizou et al. (2020) who investigated the impact of room acoustic conditions on voice performance.

3.3 *Perceptual Aspects of Listeners*

One of the key questions regarding the influence of room acoustics on live performance is whether listeners are able to perceive the adjustments implemented by musicians to accommodate the acoustics of the room. Although only a few studies have been conducted, the preliminary results suggest that listeners can at least

partially perceive the performance adjustments. Ueno et al. (2010) conducted an experiment with six listeners comparing recordings in three different rooms, and listeners were asked to judge the similarity and to provide a free description of the differences. In the 54 and 44% of the judged recordings listeners were able to perceive clear and subtle differences, respectively. Amengual Garí (2017) conducted an online test with 24 subjects comparing four different versions (recorded in different virtual rooms) of three trumpet pieces. The results suggest that listeners are able to perceive the overall sound level and timbral changes to a great extent. However, other parameters such as dynamic or tempo variations are likely multidimensional aspects e.g., dynamic variations are the result of level variations over time—and thus each listener could potentially have a more individual internal representation of those than unidimensional aspects such as overall sound level.

4 Discussion and Outlook

4.1 *Naturalness of Auralization Methods*

The most wanted feature of a virtual environment on musicians performing in it would be the lack of any artifact or unnatural perception. Given the fact that most auralization environments are built in anechoic chambers or force musicians to wear at least headphones, this feature is hard to provide. As challenging as auralization is the simulation of an adequate visual experience of musical performance. Once musicians are willing to accept shortcomings of the auralization conditions concerning a visual context or the need to wear technical equipment, a high degree of naturalness of a purely acoustical environment can be realized today. Examples are the environments provided using the SDM technique as implemented for investigation of acoustic feedback on the performance of musicians. However, the implementation of details of the acoustic environment such as reflections from music stands (Amengual and Kob, 2017), acoustic changes induced by movements of the musicians during performance (Ackermann et al., 2019) or the interaction with other musicians are beyond the current scope of acoustic simulations.

Shortcomings of current implementations of virtual performance rooms are still the interfaces between the instrument played by the musician and the perceived sound field: The immediate perception of the instrument's vibration and direct sound need to match the sound processed through the virtual environment. Even fast algorithms on high-performance computers will exhibit delays between these sound transmission paths that might be perceived by the musicians.

Another challenge is the difficulty to test and quantify the realism of virtual acoustic environments. Whereas a performance in a real acoustic environment would be a kind of *gold standard*, most set-ups would only optimize the acoustic conditions, whereas other boundary conditions such as the visual and atmospheric environment

are not virtualized. The consequence of this auditory-visual incongruence for the reliability of performance research is difficult to rate.

4.2 *Classification of Musicians*

An outcome of the studies of musicians' performance in variable acoustic environments is a rather individual reaction: Some musicians tend to increase the tempo with increased reverberation, some slow down. Some players are quite sensitive to acoustic conditions, whereas others keep their style constant disregard the acoustic environment. However, these characteristics seem to be held constant for each individual. Moreover, there seem to be certain categories of reaction patterns that musicians might be clustered into. In addition, these individual adjustments seem to be affected as well by musical character of the interpreted pieces. This clearly calls for further research in order to evaluate which aspects of music performance influence such a clustering.

4.3 *Further Investigations*

A challenge for future research could be the implementation of artifact-free virtual environments with less invasive technical boundary conditions for the musicians. One of the problems to be solved is the presence of direct sound due to structure-borne sound transmission from the instrument to the musicians' ears. Due to the need for numerical calculations of the virtual environment, a delay between the direct sound and the auralized sound reduces the naturalness of the generated sound field.

Acknowledgements A part of the methods and results presented in this chapter were obtained during a Ph.D. thesis in the frame of the Marie Curie Integrated Training Network "BATWOMAN". Further results reported here were obtained during another Ph.D. work funded by the German National Academic Foundation and the German Research Foundation (DFG WE 4057/9-1). The authors thank the musicians for their participation in the studies and Tapio Lokki for the helpful comments on the manuscript. They are further indebted to two anonymous reviewers for constructive comments.

References

- Ackermann, D., C. Böhm, F. Brinkmann, and S. Weinzierl. 2019. The acoustical effect of musicians' movements during musical performances. *Acta Acustica united with Acustica*. 105 (2): 356–367. <https://doi.org/10.3813/AAA.919319>.
- Allen, J.B., and D.A. Berkley. 1979. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America* 65 (4): 943–950. <https://doi.org/10.1121/1.382599>.

- Amengual Garí, S.V. 2017. Investigations on the influence of acoustics on live music performance using virtual acoustic methods. Ph.D. thesis, Detmold University of Music. <https://opus.hfm-detmold.de/frontdoor/index/index/searchtype/latest/docId/68/start/0/rows/10>. Accessed 07 Oct 2019
- Amengual Garí, S. V., M. Kob, and T. Lokki. 2019. Analysis of trumpet performance adjustments due to room acoustics. In *Proceedings of the International Symposium on Room Acoustics-ISRA 2019*. <http://publications.rwth-aachen.de/record/772232>.
- Amengual Gari, S. V. and M. Kob. 2017. Investigating the impact of a music stand on stage using spatial impulse responses. *Audio Engineering Society Convention* 143. <http://www.aes.org/elib/browse.cfm?elib=18622>.
- Amengual Garí, S.V., W. Lachenmayr, and M. Kob. 2015. Study on the influence of room acoustics on organ playing using room enhancement. In *Proceedings of 3rd Vienna Talk on Music Acoustics, Vienna*.
- Arend, J. M., L. Tim, and C. Pörschmann. 2019. A reactive virtual acoustic environment for interactive immersive audio. *AES International Conference on Immersive and Interactive Audio*. <http://www.aes.org/e-lib/browse.cfm?elib=20431>.
- Astolfi, A., M. Giovannini, G. Barbato, and M. Filippi. 2007. The interpretation of objective measurements on the stage by means of the correlation with subjective data. In *Proceedings of the 19th ICA, Madrid*.
- Barron, M. 1978. The gulbenkian great hall, lisbon, ii: An acoustic study of a concert hall with variable stages. *Journal of Sound and Vibration* 59 (4): 481–502.
- Barron, M.F. 1974. The effect of early reflections on subjective acoustical quality in concert halls. Ph.D. thesis, University of Southampton.
- Begault, D.R., E.M. Wenzel, and M.R. Anderson. 2001. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society* 49 (10): 904–916. <http://www.aes.org/e-lib/browse.cfm?elib=10175>.
- Beranek, L. 2004. *Concert Halls and Opera Houses—Music, Acoustics and Architecture*, 2nd ed. Berlin: Springer.
- Berkhout, A.J., D. de Vries, and J.J. Sonke. 1997. Array technology for acoustic wave field analysis in enclosures. *The Journal of the Acoustical Society of America* 102 (5): 2757–2770. <https://doi.org/10.1121/1.420330>.
- Berkhout, A.J., D. de Vries, and P. Vogel. 1993. Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America* 93 (5): 2764–2778. <https://doi.org/10.1121/1.405852>.
- Berntson, A., and J. Andersson. 2007. Investigations of stage acoustics for a symphony orchestra. In *Proceedings of the International Symposium on Room Acoustics, Sevilla*.
- Blauert, J. 2013. *The Technology of Binaural Listening*. Berlin: Springer.
- Blum, D. 1987. *The Art of Quartet Playing. The Guarneri Quartet in Conversation with David Blum*. Ithaca: Cornell Universtiy Press.
- Bolzinger, S., and J.C. Risset. 1992. A preliminary study on the influence of room acoustics on piano performance. *Journal de Physique IV* 2 (C1): 93–96. <https://doi.org/10.1051/jp4:1992116>.
- Bolzinger, S., O. Warusfel, and E. Kahle. 1994. A study of the influence of room acoustics on piano performance. *Journal de Physique IV* 4 (C5): 617–620.
- Borciani, P. 1973. *Das Streichquartett*. Mailand: Ricordi.
- Borish, J. 1984. Extension of the image model to arbitrary polyhedra. *Journal of the Acoustical Society of America* 75 (6): 1827. <https://doi.org/10.1121/1.390983>.
- Botteldooren, B. 1995. Finite-difference time-domain simulation of low-frequency room acoustic problems. *Journal of the Acoustical Society of America* 98 (6): 3302–3308.
- Brereton, J.S., D.T. Murphy, and D.M. Howard. 2012. The virtual singing studio: A loudspeaker-based room acoustics simulation for real-time musical performance. In *Proceedings of the Joint Baltic-Nordic Acoustics Meeting*, 8.

- Brinkmann, F., and A. Lindau. 2012. Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings. *Journal of the Audio Engineering Society* 60 (1/2): 54–62.
- Brinkmann, F., A. Lindau, and S. Weinzierl. 2017. On the authenticity of individual dynamic binaural synthesis. *The Journal of the Acoustical Society of America*. 142, 1784. <https://doi.org/10.1121/1.5005606>.
- Brinkmann, F., L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl. 2019. A round robin on room acoustical simulation and auralization. *The Journal of the Acoustical Society of America*. 145, 2746. <https://doi.org/10.1121/1.5096178>.
- Brunskog, J., A.C. Gade, G.P. Bellester, and L.R. Calbo. 2009. Increase in voice level and speaker comfort in lecture rooms. *Journal of the Acoustical Society of America* 125 (4): 2072–2082. <https://doi.org/10.1121/1.3081396>.
- Cannam, C., C. Landone, and M. Sandler. 2010. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, 1467—1468. <https://doi.org/10.1145/1873951.1874248>.
- Chiang, W., S. Chen, and C. Huang. 2003. Subjective assessment of stage acoustics for solo and chamber music performances. *Acta Acustica/Acustica* 89: 848–856.
- Czerny, C. 1839. *Vollständige theoretisch-practische Pianoforte-Schule* [Complete theoretical-practical piano school], vol. 3. Wien: Diabelli.
- Dammerud, J. 2009. Stage acoustics for symphony orchestras in concert halls. Doctoral thesis, University of Bath.
- Dammerud, J., M. Barron, and E. Kahle. 2010. *Objective assessment of acoustic conditions on concert hall stages—limitations and new strategies*. Melbourne: In Proceeding of the ISRA.
- Dammerud, J., M. Barron, and E. Kahle. 2011. Objective assessment of acoustic conditions for symphony orchestras. *Building Acoustics* 18 (3–4): 207–219. <https://doi.org/10.1260/1351-010X.18.3-4.207>.
- Daniel, J. 2003. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. In *Audio Engineering Society Conference: 23rd International Conference: Signal Processing in Audio Recording and Reproduction*. <http://www.aes.org/e-lib/browse.cfm?elib=12321>.
- Devaney, J. 2016. Inter- versus intra-singer similarity and variation in vocal performances. *Journal of New Music Research* 45 (3): 252–264. <https://doi.org/10.1080/09298215.2016.1205631>.
- Downie, J.S. 2003. Music information retrieval. *Annual Review of Information Science and Technology* 37: 295–340. <https://doi.org/10.1002/aris.1440370108>.
- Eerola, T., and P. Toivainen. 2004. MIR in Matlab: The MIDI toolbox. In *Proceedings of the International Symposium on Music Information Retrieval Conference (ISMIR)*, Barcelona, Spain. <http://ismir2004.ismir.net/proceedings/p004-page-22-paper193.pdf>. Accessed 9 Oct 2019.
- Flesch, C. 1928. *Die Kunst des Violinspiels (The art of playing the violin), II: Künstlerische Gestaltung und Unterricht (Artistic design and education)*. Berlin: Verlag Ries & Erler.
- Friberg, A., E. Schoonderwaldt, and A. Hedblad. 2011. Perceptual ratings of musical parameters. In *Gemessene Interpretation—Computergestützte Aufführungsanalyse im Kreuzverhör der Disziplinen*, eds. H. Loesch and S. Weinzierl (Schott (Klang und Begriff 4), Chap. Perceptual, 237–253. <http://kth.diva-portal.org/smash/record.jsf?pid=diva2:465496>. Accessed 9 Oct 2019.
- Friberg, A., E. Schoonderwaldt, and P.N. Juslin. 2007. CUEx: An algorithm for automatic extraction of expressive tone parameters in music performance from acoustic signals. *Acta Acustica united with Acustica* 93: 411–420.
- Gabriellson, A. 1999. The performance of music. In *The Psychology of Music*, ed. D. Deutsch, 2nd ed., 501–602. New York: Academic Press.
- Gade, A.C. 1986. Acoustics of the orchestra platform from the musicians' point of view. In *Acoustics for Choir and Orchestra*, ed. S. Ternström, vol. 52 (Royal Swed. Acad. of Music), 23–42.
- Gade, A.C. 1989a. Investigations of musicians' room acoustic conditions in concert halls. Part I: Methods and laboratory experiments. *Acustica* 69: 193–203.

- Gade, A.C. 1989b. Investigations of musicians' room acoustic conditions in concert halls. Part II: Field experiments and synthesis of results. *Acustica* 69: 249–262.
- Gade, A.C. 1992. Practical aspects of room acoustic measurements on orchestra platforms. In *Proceeding of the 14th ICA*, Beijing.
- Gade, A.C. 2010. *Acoustics for symphony orchestras: Status after three decades of experimental research*. Melbourne: In Proceedings of the International Symposium on Room Acoustics (ISRA), Melbourne August 29–31, 2010.
- Gade, A.C. 2013. *Subjective and objective measures of relevance for the description of acoustics conditions on orchestra stages*. Toronto: In Proceeding of International Symposium on Room Acoustics (ISRA), Toronto June 9–11, 2013.
- Galamian, I. 1962. *Principles of Violin Playing and Teaching*. Englewood Cliffs: Prentice-Hall.
- Gardner, W.G. 1995. Efficient convolution without input-output delay. *Journal of the Audio Engineering Society* 43 (3): 127–136.
- Gerzon, M.A. 1973. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society* 21 (1): 2–8.
- Gingras, B. 2014. Perceiving musical individuality: Introduction to the research topic. *Frontiers in Psychology* 5 (661). <https://doi.org/10.3389/fpsyg.2014.00661>.
- Gingras, B., P.-Y. Asselin, S. McAdams. 2013. Individuality in harpsichord performance: Disentangling performer- and piece-specific influences on interpretive choices. *Frontiers in Psychology* 4 (895). <https://doi.org/10.3389/fpsyg.2013.00895>.
- Goebel, W., S. Dixon, G. De Poli, A. Friberg, R. Bresin, and G. Widmer. 2008. Sense in expressive music performance: Data acquisition, computational studies, and models. In *Sound to Sense, Sense to Sound—A State of the Art in Sound and Music Computing*, ed. P. Polotti, and D. Rocchesso. Berlin: Logos Verlag.
- Griesinger, D. 1997. The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces. *Acta Acustica united with Acustica* 83 (4): 721–731.
- Guthrie, A., S. Clapp, J. Braasch, and N. Xiang. 2013. Using ambisonics for stage acoustics research. In *Proceeding of International Symposium on Room Acoustics*, 1–10. https://pdfs.semanticscholar.org/9286/57d91ec3ca5258c4882e8acedc0cb5c2ab6f.pdf?_ga=2.51235498.2008809562.1570482125-647436850.1570482125. Accessed 07 Oct 2019.
- ISO 3382–1. 2009. *Acoustics—Measurement of Room Acoustic Parameters—Part 1: Performance Spaces*. Geneva: International Organization for Standardization.
- Jeon, J.Y., Y.S. Kim, H. Lim, and D. Cabrera. 2015. Preferred positions for solo, duet, and quartet performers on stage in concert halls: In situ experiment with acoustic measurements. *Building and Environment* 93 (Part 2): 267–277. <https://doi.org/10.1016/j.buildenv.2015.07.010>.
- Kato, K., K. Ueno, and K. Kawai. 2007. Musicians' adjustments of performance to room acoustics. part ii: Acoustical analysis of performed sound signals. in *Proceeding of the 19th ICA*, Madrid.
- Kato, K., K. Ueno, and K. Kawai. 2015. Effect of room acoustics on musicians' performance. part ii: Acoustic analysis of the variations in performed sound signals. *Acta Acustica united with Acustica* 101 (4): 743–759. <https://doi.org/10.3813/AAA.918870>.
- Kleiner, M., B. Dalenbäck, and P. Svensson. 1993. Auralization an overview. *Journal of the Audio Engineering Society* 41 (11): 861–875. <http://www.aes.org/e-lib/browse.cfm?elib=6976>.
- Kuttruff, H. 2009. *Room Acoustics*, 5th ed. Didcot: Taylor & Francis.
- Laird, I., D.T. Murphy, P. Chapman, and J. Seb. 2011. Development of a virtual performance studio with application of virtual acoustic recording methods. In *130th Convention of the Audio Engineering Society*, New York, 12. Preprint 8358, <http://www.aes.org/e-lib/browse.cfm?elib=1582>.
- Lartillot, O., and P. Toiviainen. 2007. A Matlab toolbox for musical feature extraction from audio. In *Proceeding of 10th International Conference on Digital Audio Effects (DAFx-07)*, Bordeaux. <https://dafx.labri.fr/main/papers/p237.pdf>. Accessed 07 Oct 2019.
- Leman, M. 2008. *Embodied Music Cognition and Mediation Technology*. Cambridge MA/London: MIT Press.
- Lentz, T. 2006. Dynamic cross-talk cancellation for binaural synthesis in virtual reality environments. *Journal of the Audio Engineering Society* 54 (4): 283–294.

- Lerch, A. 2008. Software-based extraction of objective parameters from music performances. Ph.D. thesis, TU Berlin.
- Luizard, P., J. Steffens, and S. Weinzierl. 2020. Singing in different rooms: Common or individual adaptation patterns. *The Journal of the Acoustical Society of America*. 147 (2): EL132–EL137. <https://doi.org/10.1121/10.0000715>.
- Malham, D.G., and A. Myatt. 1995. 3-d sound spatialization using ambisonic techniques. *Computer Music Journal* 19 (4): 58–70.
- Marshall, A.H., D. Gottlob, and H. Alrutz. 1978. Acoustical conditions preferred for ensemble. *Journal of the Acoustical Society of America* 64 (5): 1437–1442. <https://doi.org/10.1121/1.382121>.
- Marshall, A.H., and J. Meyer. 1985. The directivity and auditory impressions of singers. *Acustica* 58: 130–140.
- McAdams, S., P. Depalle, and E. Clarke. 2004. Analyzing musical sound. In *Empirical Musicology*, ed. E. Clarke, and N. Cook, 157–196. Oxford: Oxford University Press.
- McKay, C., and I. Fujinaga. 2009. jMIR: Tools for automatic music classification. In *Proceedings of the International Computer Music Conference (ICMC '09)*, 65–68. <https://www.semanticscholar.org/paper/jMIR%3A-Tools-for-Automatic-Music-Classification-McKay-Fujinaga/a8351d9f9a9c3705ab98219d706b4668da4ac376>. Accessed 07 Oct 2019.
- Meyer, J., and E.C. Biassoni de Serra. 1980. Zum Verdeckungseffekt bei Instrumentalmusikern [On masking with instrumental musicians]. *Acta Acustica/Acustica* 46 (2): 130–140.
- MIDI Manufacturer's Association (MMA). 1996. The complete midi 1.0 detailed specification. midi.org. 2020. Details about MIDI 2.0^T M, MIDI-CI, Profiles and Property Exchange. <https://www.midi.org/articles-old/details-about-midi-2-0-midi-ci-profiles-and-property-exchange>. Accessed 07 Oct 2019.
- Møller, H., M.F. Sørensen, C.B. Jensen, and D. Hammershøi. 1996. Binaural technique: Do we need individual recordings?. *Journal of the Audio Engineering Society* 44 (6): 451–469. <http://www.aes.org/e-lib/browse.cfm?elib=7897>.
- Nakamura, T. 1987. The communication of dynamics between musicians and listeners through musical performance. *Perception and Psychophysics* 41 (6): 525–533.
- Naylor, G.M. 1988. Modulation transfer and ensemble music performance. *Acustica* 65: 127–137.
- O'Keefe, J. 1995. Acoustic conditions in orchestra pits and proscenium arch theatres. In *Proceeding of the Institute of Acoustics*, vol. 17, 133.
- Ondet, A.M., and J.L. Barbry. 1989. Modeling of sound propagation in fitted workshops using ray tracing. *Journal of the Acoustical Society of America* 85 (2): 787–796.
- Panton, L., D. Cabrera, and D. Holloway. 2016. Using a spherical microphone array for stage acoustics: A preliminary case for a new spatial parameter. In *Proceedings of the 22nd International Congress on Acoustics*, Buenos Aires.
- Panton, L., D. Holloway, D. Cabrera, and L. Miranda. 2017. Stage acoustics in eight australian concert halls: Acoustic conditions in relation to subjective assessments by a touring chamber orchestra. *Acoustics Australia* 45 (1): 25–39. <https://doi.org/10.1007/s40857-016-0075-2>.
- Pietrzyk, A., and M. Kleiner. 1997. The application of the finite element method to the prediction of sound fields of small rooms at low frequencies. In *102th Convention of the Audio Engineering Society*, Gothenburg.
- Pulkki, V. 1997. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society* 45 (6): 456–466.
- Pulkki, V. 2006. Directional audio coding in spatial sound reproduction and stereo upmixing. In *Audio Engineering Society Conference: 28th International Conference: The Future of Audio Technology—Surround and Beyond*. <http://www.aes.org/e-lib/browse.cfm?elib=13847>.
- Quantz, J.J. 1752. *Versuch einer Anweisung, die Flöte traversière zu spielen* [Attempted instruction to play the transverse flute]. Kassel: Bärenreiter (1983).
- Repp, B.H. 1994. On determining the basic tempo of an expressive music performance. *Psychology of Music* 22 (2): 157–167.
- Repp, B.H. 1999. Effect of auditory feedback deprivation on expressive piano performance. *Music Perception* 16 (4): 409–438.

- Sanders, J. 2003. Suitability of New Zealand halls for chamber music. Marshall Day Acoustics. <http://marshallday.com>. Accessed 07 Oct 2019.
- Sarvazyan, A.P., M.W. Urban, and J.F. Greenleaf. 2013. Acoustic waves in medical imaging and diagnostics. *Ultrasound in Medicine and Biology* 39 (7): 1133–1146.
- Savioja, L., and U.P. Svensson. 2015. Overview of geometrical room acoustic modeling techniques. *Journal of the Acoustical Society of America* 138 (2): 708–730. <https://doi.org/10.1121/1.4926438>.
- Schärer, Z., and A. Lindau. 2009. Evaluation of equalization methods for binaural signals. In *Proceeding of the 126th Convention of the Audio Engineering Society*, Preprint 7721, Munich.
- Schärer Kalkandjiev, Z. 2015. The influence of room acoustics on solo music performances. An empirical investigation. Ph.D. thesis, TU Berlin. <https://doi.org/10.14279/depositonce-4785>.
- Schärer Kalkandjiev, Z., and S. Weinzierl. 2013. The influence of room acoustics on solo music performance: An empirical case study. *Acta Acustica united with Acustica* 99 (3): 433–441. <https://doi.org/10.3813/AAA.918624>.
- Schärer Kalkandjiev, Z., and S. Weinzierl. 2015. The influence of room acoustics on solo music performance. an experimental study. *Psychomusicology: Music, Mind, and Brain* 25 (3): 195–207. <https://doi.org/10.1037/pmu0000065>.
- Seashore, C.E. 1938. *Psychology of Music*. New York: Dover Publications.
- Sloboda, J.A. 1982. Music performance. In *The Psychology of Music*, 1st ed, ed. D. Deutsch, 479–496. New York: Academic Press.
- Sloboda, J.A. 2000. Individual differences in music performance. *Trends in Cognitive Sciences* 4 (10): 397–403. [https://doi.org/10.1016/S1364-6613\(00\)01531-X](https://doi.org/10.1016/S1364-6613(00)01531-X).
- Spohr, L. 1833. *Violinschule* [violin school]. Wien: Haslinger.
- Tervo, S., J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki. 2015. Spatial analysis and synthesis of car audio system and car-cabin acoustics with a compact microphone array. *Journal of the Audio Engineering Society* 63 (11): 914–925.
- Tervo, S., J. Pätynen, and T. Lokki. 2013. Spatial decomposition method for room impulse responses. *Journal of the Audio Engineering Society* 61 (1): 1–13.
- Ueno, K., T. Kanamori, and H. Tachibana. 2005. Experimental study on stage acoustics for ensemble performance in chamber music. *Acoustical Science and Technology* 4 (4): 345–352.
- Ueno, K., K. Kato, and K. Kawai. 2007. Musicians' adjustments of performance to room acoustics. Part I: Experimental performance and interview in simulated sound field. In *Proceeding of the 19th International Congress on Acoustics, 1807–1812*. Madrid.
- Ueno, K., K. Kato, and K. Kawai. 2010. Effect of room acoustics on musicians' performance. Part I: Experimental investigation with a conceptual model. *Acta Acustica united with Acustica* 96: 505–515. <https://doi.org/10.3813/AAA.918303>.
- Ueno, K., and H. Tachibana. 2003. Experimental study on the evaluation of stage acoustics by musicians using a 6-channel sound simulation system. *Acoustical Science and Technology* 24 (3): 130–138.
- Ueno, K., K. Yasuda, H. Tachibana, and T. Ono. 2001. Sound field simulation for stage acoustics using 6-channel system. *Acoustical Science and Technology* 22 (4): 307–309. <https://doi.org/10.1250/ast.22.307>.
- van den Braak, E.W., and R.C. van Luxemburg. 2008. New (stage) parameter for conductor's acoustics? In *Proceeding of Acoustics '08*, Paris, 2145–2150.
- van Luxemburg, R.C., C.C. Hak, P.H. Heijnen, and M. Kivits. 2009. Stage acoustics: Experiments on 7 stages of concert halls in the Netherlands. In *Proceeding of inter-noise 2009*, Ottawa.
- Van Vugt, F.T., H.-C. Jabusch, and E. Altenmüller. 2013. Individuality that is unheard of: Systematic temporal deviations in scale playing leave an inaudible pianistic fingerprint. *Frontiers in Psychology* 4 (134). <https://doi.org/10.3389/fpsyg.2013.00134>.
- von Békésy, G. 1968. Feedback phenomena between the stringed instrument and the musician. *The Rockefeller University Review* 6 (2):

- Weinzierl, S., S. Lep, and M. Thiering. 2020. The language of rooms: From perception to cognition. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 435–454. Cham, Switzerland: Springer and ASA press.
- Weinzierl, S., and M. Vorländer. 2015. Room acoustical parameters as predictors of room acoustical impression: What do we know and what would we like to know? *Acoustics Australia* 1–8.
- Wenmaekers, R., C. Hak, M. Hornikx, and A. Kohlrausch. 2017. Sensitivity of stage acoustic parameters to source and receiver directivity: Measurements on three stages and in two orchestra pits. *Applied Acoustics* 123 (Supplement C): 20–28. <https://doi.org/10.1016/j.apacoust.2017.03.004>.
- Wenmaekers, R.H., C.C. Hak, and L.C. van Luxemburg. 2012. On measurements of stage acoustic parameters: Time interval limits and various source-receiver distances. *Acta Acustica/Acustica* 98: 776–789.
- Wenzel, E.M., M. Arruda, D.J. Kistler, and F.L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America* 94 (1): 111–123. <https://doi.org/10.1121/1.407089>.
- Winckel, F.F. 1962. Optimum acoustic criteria of concert halls for the performance of classical music. *Journal of the Acoustical Society of America* 34 (1): 81–86.

Binaural Modeling from an Evolving-Habitat Perspective



Jonas Braasch

Abstract Functional binaural models have been used since the mid-20th century to simulate laboratory experiments. The goal of this chapter is to extend the capabilities of a cross-correlation model so it can demonstrate human listening in complex scenarios found in nature and human-built environments. A ray-tracing model is introduced that simulates a number of environments for this study. This chapter discusses how the auditory system is used to read and understand the environment and how tasks that require binaural hearing may have evolved throughout human history. As use cases, sound localization in a forest is examined, as well as the binaural analysis of spatially diffuse and rectangular rooms. The model is also used to simulate binaural hearing during a walk-through a simulated office-suite environment.

1 Introduction

The goal of this chapter is to examine binaural models from an evolving-habitat perspective. While the evolution of the auditory system has been studied extensively from a phylogenetic perspective to establish knowledge of how the auditory system developed anatomically over time, the auditory system's ability to adapt to changing habitats over tens of thousands of years has been hardly investigated. Since it is impossible to travel back in time, the topic cannot be studied directly. This chapter describes an attempt of an initial study examining this by simulating different environments with a ray tracing model and using an extended binaural model for an auditory-specific analysis. When studying how the auditory system can adapt to different habitats, one must keep in mind that the anatomical changes of the auditory system took place over millions of years. While the structure of the auditory system continues to change over time, these changes occur at a much slower pace than most sociological changes. Thus, it can be assumed that our auditory system is basically

J. Braasch (✉)
School of Architecture, Rensselaer Polytechnic Institute,
Troy, NY 12180, USA
e-mail: braasj@rpi.edu

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_10

structured the same way that it was during the beginning of modern civilization, which started about 50,000 years ago (Peck 1994). Since the neurological structure of the brain is very flexible, mammals and other organisms can easily adapt to new environments and situations—especially during the early post-natal phase (Peck 1995). This flexibility allows us to adjust to new sonic environments. Humans can, for example, comprehend and appreciate classical music in modern concert halls using an auditory system that primarily developed in natural habitats.

Traditionally, binaural models have been designed to simulate laboratory scenarios, for example, to predict the lateral position of a binaural stimulus presented over headphones. In this chapter, it is attempted to extend this knowledge for better understanding and predicting binaural-hearing tasks in natural environments and other complex situations that arose as civilization evolved and the built environment changed. Also, a bridge will be created to the robotic community, which has its own distinct way of designing sound-sensing systems. Experts in robotics often attempt to solve tasks in complex environments, for example, acoustically navigating systems, but without the desire to understand how these tasks are accomplished in biological systems. In the context of this chapter, understanding will be defined as *the ability to make judgments from perceived information*. In some cases, the understanding consists of the ability to accurately decode the intended meaning sent by a communication partner, for example, a conversation partner, or the ability to interpret unintended cues—such as the sounds of an approaching predator. In any case, understanding allows us to infer something from the received acoustic signals, and these signals then become information.

Evolutionary biologists agree that a biological organism needs to be successful in this behavioral complex of four tasks to survive as a species, namely, (i), *feeding*, (ii), *fleeing*, (iii), *fighting*, and (iv), *flirting (reproduction)*¹—compare, for instance, Graham (2014). Spatial awareness is essential to success in all these goals—to find food, avoid predators and to communicate with tribe members for various reasons ranging from cooperation to mating. As a starting point for the binaural analysis, the need for spatial acoustic communication and sensing will now be examined in the view of the main four survival tasks mentioned above.

1.1 Feeding

The early *Homo Sapiens* survived mainly as hunters and gatherers. Unlike other vertebrates, such as barn owls or bats, who find their prey acoustically, humans localize prey or gather objects using vision as their primary sense. Consequently, the acoustic-localization performance does not need to be as accurate as is the case for acoustical hunters, who must target their prey precisely. Most likely, acoustic communication between tribe members played a big role when hunting animals, for example, when engaging in an attack. Studies have found evidence that the early homo

¹Also known as the four F's: feeding, fleeing, fighting, fornication.

sapiens lived in the plains and hunted large animals from a distance using spears and other long-distance weapons—Villa and Soriano (2010). The ability to follow these hunting patterns was a direct result of the *Cognitive Revolution*—Wynn and Coolidge (2004, 2008), Coolidge and Wynn (2018). The rise of new cognitive abilities enabled homo sapiens to plan ahead, conduct better-coordinated group hunting and also to spatially navigate larger terrains—Baril (2012).

In contrast, *Homo Neanderthalensis* is believed to have been a hunter who killed animals in close combat, based on the type of spear perforations found in deer skeletons—Gaudzinski-Windheuser et al. (2018) and other evidence. In this context, it is noteworthy that homo sapiens has a voice box that is very different from that of *Homo Neanderthalensis* and other early human species. This results in a lower fundamental pitch (Fitch 2000). The need for lowering the voice could have resulted from the need to communicate acoustically over larger distances in the plain as the homo sapiens started to specialize in hunting animals from a distance. This would also explain why the frequency range is lower than it is the case of many other mammals. The first mammals were small nocturnal animals who presumably lived in densely vegetated areas—Gerkema et al. (2013). Their high-frequency hearing range is optimal for localizing potential predators at a close distance (Joris and Trussell 2018). However, high frequencies are not optimal for localizing sound sources from a larger distance because of air absorption (dissipation), which increases with frequency.

1.2 Fleeing

In contrast to when hunting prey, the angular localization accuracy is not that critical when fleeing from predators because one usually runs away from them. However, it is essential to detect the predators early on before they pose an imminent danger. Auditory cues always become predominant when visual cues are not available. This is the case when it is too dark to see, visual objects are occluded, or the acoustic sources are outside the visual field. The auditory sense also monitors the environment during sleep, and it can be shown that children are not yet disturbed by sounds at night (Busby et al. 1994). One explanation for this observation is that, from an evolutionary perspective, it is better for children at certain ages to have an undisturbed sleep and rely on their parents for monitoring than to monitor the environment themselves. A study on detecting fire alarms revealed that the sleep of children is often so deep that children do not wake up when the alarm is set off (Bruck 1999).

In discussions it is usually emphasized that the ability to detect signals to monitor predators is of particular relevance. However, an absence of sound can be equally important, because other animals will quiet down in a sector from where a predator is approaching. This might be one of the reasons for enjoying immersive sounds, namely, that they can serve as an inherent indicator that no predator is approaching.² While this hypothesis remains to be proven, several studies have shown that human subjects

²Personal communication with David Mountain, Boston University, April 5, 2013.

do not feel comfortable when performing tasks in extremely quiet environments (Volf 2012; DeLoach et al. 2015). Alternative strategies are pursued where fleeing is not an option. The females of the indigenous *BaYaka* group, for example, gather together and shouted group calls to make themselves appear as a large and well-coordinated group to scare away predators in lieu of fleeing (Knight and Lewis 2017, p. 442).

1.3 Fighting

Combat between humans is as nearly as old as the homo sapiens (Meller and Schefzik 2015; Ferrill 2018), and the remains of the oldest homicide victim found are 430,000 years old (Sala et al. 2015). The acoustic requirements for human combat against other humans or predators are very similar to the acoustic communication in hunting situations discussed above. Shouting calls are essential to coordinate attacks and warn others from counter attacks, demanding excellent spatial-hearing skills.

1.4 Reproduction

Many animals mostly rely on acoustic signaling and sound localization to find mating partners. Although many animals shout out mating calls over long distances, it is more likely human courting has always been an intimate social interaction, since humans always lived together in groups. It is widely believed that the fundamental pitch differences in human female and male voices have evolved to make each other more attractive to the opposite sex (Jones et al. 2010). In the context of spatial hearing, it is interesting though that homo sapiens engaged in artistic activities from early on that appear to address both erotic and musical desires. For example, in the Hohle-Fels cave, bone flutes were found next to venus figurines and phallus sculptures (Conard and Wolf 2014; Conard et al. 2009). Since then, the flute has been a typical courting instrument in indigenous cultures, see for example Conlon (2004). Adjacent to the inhabited part of the Hohle-Fels cave, where the bone flute was excavated, a much larger cavern exists with a reverberation time of about 2 s. It is not hard to imagine that our ancestors would have played the flutes in this larger cavern to enjoy the acoustics. At least, it is known that early humans were very aware of their acoustic environment. For instance, Reznikoff found that many prehistoric cave drawings were painted at places with dominant acoustic resonances (Reznikoff 2004/2005).

1.5 Modality and Bandwidth

In order to understand how our auditory system developed and was utilized, one needs to examine what cues and mechanisms are available to process these cues.

This way, situations can be determined in which acoustic cues supersede other cues. Of the five major senses, touch, taste, smell, vision, and audition, only the last three are useful to sense objects from a distance. Our olfactory sense is less developed than that of other species, including dogs. Human beings only have a directional sense of smell when they are moving, and even though it can be useful to detect the presence of a predator or food, this sense lacks the spatial precision of the auditory and visual systems. The early homo sapiens was primarily a diurnal hunter using the visual sense to hunt animals from a distance with spears and to collect food from plants. Most likely, the auditory sense had initially a support role until speech communication became increasingly important. The visual sense is limited to the binocular visual field and covers only about 214° in the horizontal plane (Rönne 1915). In contrast, the auditory sense is not spatially restricted, and it also helps us monitor our environment at night. Sound localization has been important to detect the direction of a predator quickly. It can also be assumed that it was important for our ancestors to localize each others' voice commands when tracking prey. Sound localization has also always been important in environments where vision is partially obstructed, for example when hunting deer in a dense forest (Gaudzinski-Windheuser et al. 2018; Roebroeks et al. 1992). Our first example deals with such a situation. A virtual walk-through in a forest is presented and analyzed in the next section.

2 Simulating Sound Localization in a Forest with Partially Obstructed Sight Using a Ray-Tracing Model

2.1 Introduction

To be able to better understand how the binaural system evolved over millions of years to perform robustly in complex environments, several scenarios were developed in which auditory cues are particularly important. Obviously, sound localization is always in demand when the source is out of sight. Aside from monitoring the environment at night, forests are a good test case because trees and other vegetation typically visually obstruct objects, and some of these objects might be looking for dinner. The forest simulation was set up using a ray-tracing simulation program, which is described in the next section. Circular boundaries are used as acoustic objects to simulate the acoustic behavior of tree trunks. The forest is simulated by randomly creating circles in an area of $100 \times 100 \text{ m}^2$. The diameters of the tree trunks are set by a stochastic process.

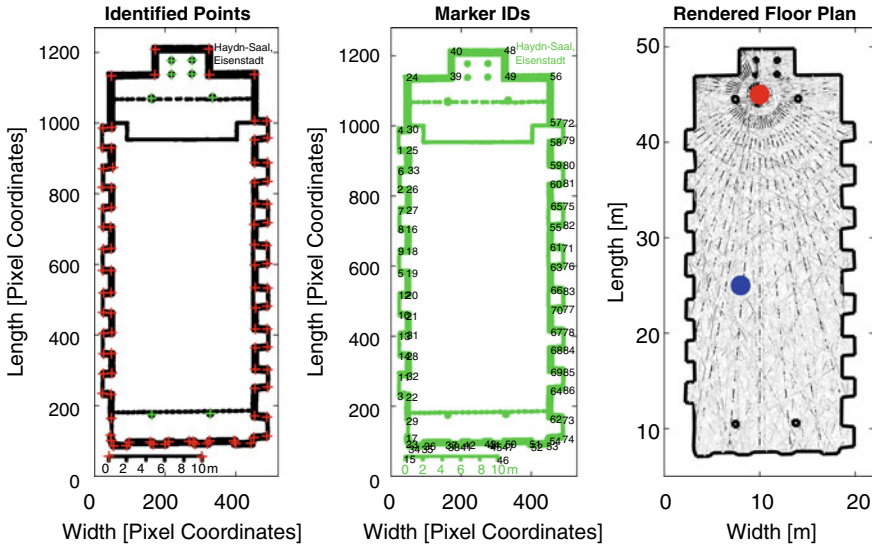


Fig. 1 Demonstration of the ray-tracing algorithm for the Haydn-Saal in Eisenstadt, Austria, where Joseph Haydn was active. **(Left)** Processed marked-up floor plan for visual inspection. **(Center)** Assignment of wall-corner identification numbers. **(Right)** Geometric model with sound source (red dot), receiver (blue dot), and calculated rays. The original floor plan was obtained from Meyer (1978, p. 147)

2.2 Creating Geometric Models

The method presented here is confined to simulations in the horizontal plane (two-dimensional rendering method) to allow fast calculations. This enables the simulation of complete walk-throughs using a batch process.³ The ray-tracing implementation was programmed in Matlab following common practice—details in Vorländer (1989), Lehnert and Blauert (1992a, b), Blauert et al. (2000). Additional features were added where needed, for example, an algorithm to create circular boundaries to simulate tree trunks and a method to generate models from floor plans rapidly.

Coordinates of acoustic boundaries can be assigned to the ray-tracing algorithm in three different ways, (i) line segments with start and end points representing walls, (ii) squared pillars with the center coordinates and the pillar width, (iii) circles represented by center and radius coordinates. Figure 1 shows an example of the ray-tracing software for a concert hall in Eisenstadt, Austria, that was recreated from a floor plan. The software can work with annotated floor plans. For this purpose floor plans are marked up within a standard bitmap editor (GIMP, Photoshop, etc.) using red dots for room corners, green dots for squared pillars, and blue dots for circles—see Fig. 1, left graph. In the next step, the program plots annotated points on top of

³Originally, the ray-tracing method was implemented to create auralizations for the horizontal array of 128-channel loudspeakers at Rensselaer’s CRAIVE-Lab.

the map providing a unique index for each annotated point—see Fig. 1, center graph. The user then creates a list of how the red points connect to walls. The scale of the floor plan needs to be annotated, and the original dimensions need to be handed over to the program as well (e.g., 1 m and 10 m for two scale points respectively). The program then transforms these data points to an editable list of geometrical objects that can be extended by the user, for example, by adding identifiers for wall materials.

The Ray-Tracing Algorithm

The program sends out rays from a user-specified source position. The program computes the rays for equidistant azimuths covering the full 360° angular range. Intersection points are computed for each ray and boundary object as shown in Fig. 2, left graph. For each ray, the closest boundary intersection is determined. At the intersections, the reflection angle is calculated using Snell’s law, which predicts that the reflected angle measured from the normal of a plane surface equals the incoming angle, that is, $\cos(\alpha_o) = \cos(\alpha_i)$. Consequently, the next-order ray is sent out into the new direction until the maximum order (e.g., the number of reflections) as specified by the user is reached. The outgoing rays are stored as a sequence of ray elements containing the intersection points and the boundary-material identifiers. Since the initial angles are stored with the rays, a source-specific directivity pattern can be simulated after all rays have been traced.

Creating a Binaural Room Impulse Response

Next, the rays are collected by a receiver, which can be located anywhere in the rendered room. For this purpose, a virtual circle with an adjustable diameter is posi-

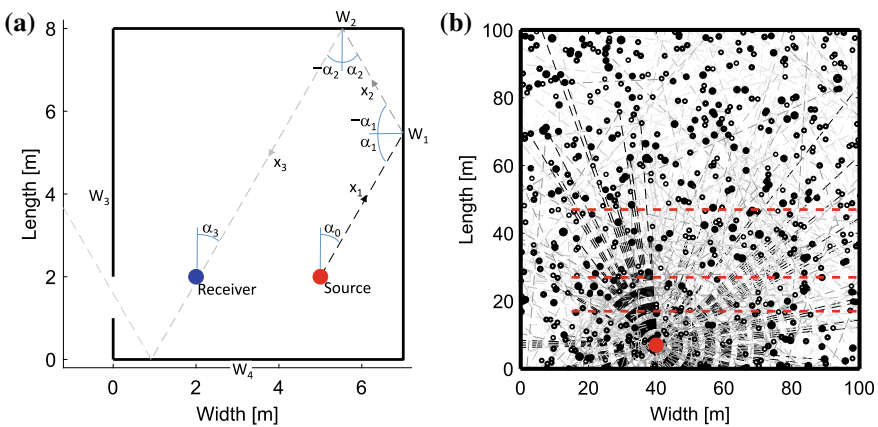


Fig. 2 (Left) Diagram to illustrate the ray-tracing method. (Right) Schematic of the simulated forest environment (top view)

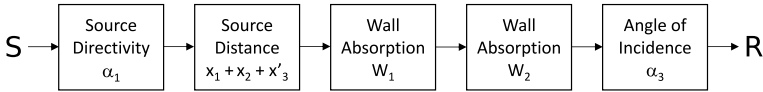


Fig. 3 Ray-tracing signal flow

tioned at the receiver location. Then, the algorithm calculates which ray elements intersect the circle, and for all positive cases, the total ray distance between sound source and receiver is calculated. All listed values are stored together with the final angle of incidence, that is, the angle that the ray was initially sent out, the reflection order, and the sequence of identifiers of the walls that the ray has hit on its way to the receiver. Based on these data, the impulse response is computed. The direct sound and the reflection are computed as delta peak at the delay that corresponds to the path length that the ray has traveled from the source to the receiver. In addition, each impulse is transformed in the following way—see also Fig. 3.

1. The magnitude of the ray is reduced based on the inverse-square law.
2. The high frequencies are filtered out based on dissipation effects in the air.
3. The absorption coefficients of the walls and other boundaries are simulated using a cascaded Finite Impulse Response (FIR) filter. These material-specific filters are chosen from a publicly available database (DIN 1968). The number of cascaded filters matches the order of the reflection.
4. In the final step, the incoming direct sound and the reflections are selected by their close proximity passing the receiver position. These are then filtered with *Head Related Transfer Functions* (HRTFs) that correspond to the closest available HRTF measurement for the direction of incidence. At this point, the room impulse response is transformed into a stereo signal. An overlap-method ensures that the delayed reflections can partially overlap.

Simulation of Late Reverberation

Late diffuse reverberation is computed in addition to the early reflections that are generated by the ray-tracing model. Since the late reverberation tail is formed by a stochastic process with an underlying Gaussian distribution, the fine structure of the simulated reverberation tail is constructed from a Gaussian noise sample. The duration of the Gaussian noise sample is adjusted to twice the value of the maximum reverberation time. Next, the noise sample is processed through a filter bank with nine adjacent octave-wide bandpass filters. An exponentially decaying time window, y_k , adjusted to the frequency-specific reverberation time, is calculated for each octave band, k :

$$y_k = e^{\frac{-t \cdot 20 \cdot \log(10)}{T_k \cdot 60}}. \quad (1)$$

with the reverberation time, T_k , in the k th frequency band, and the time, t , in seconds. Afterward, the total exponentially-decaying noise signal, x_t , is reassembled by summing up, sample-by-sample, the octave-filtered noise signals, multiplied with the exponentially decaying time window:

$$x_t = \sum_{k=1}^K x_k \cdot y_k. \quad (2)$$

This process is repeated for each channel, two for a binaural signal in each case, using independent Gaussian noise samples for each channel while keeping all other parameters constant. The frequency-specific reverberation times, T , is calculated using the Eyring formula, which is based on a three-dimensional room model that takes the room volume and the effective absorptive surface area into account, namely,

$$T_{60} = 0.161 \cdot V / (A + 4m \cdot V) \text{ s}, \quad (3)$$

whereby, A is total effective absorption, defined as the sum of all surface elements, S_k , multiplied with their specific absorption coefficients, α_k :

$$A = \left(\sum_{k=1}^K \alpha_k \cdot S_k \right). \quad (4)$$

In order to estimate the room volume, V , the area of the floor plan is calculated and then multiplied by the average room height, which has to be provided to the program. The formula for this calculation is:

$$V = A_F \cdot h. \quad (5)$$

The total effective absorption is calculated from the wall elements in the ray-tracing model, each multiplied with the average height. The frequency-specific absorption coefficients are determined with the values stored in the DIN database (DIN 1968) via the wall material identifiers. A linear onset ramp is calculated to gradually blend in the late reverberation tail with the direct sound and early reflections. The starting and end points of the ramps can be adjusted by the user.

Two methods are available to calculate the direct-to-reverberant energy ratio. The first method estimates the critical distance, the distance from the sound source at which the sound-pressure levels of the direct sound matches the sound-pressure level of the reverberant field. For an omnidirectional sound source, the critical distance can be calculated using the following equation (Kuttruff 2000, p.317),

$$r_c = 0.057 \cdot \sqrt{\frac{\gamma V}{T}}, \quad (6)$$

with the volume, V , the reverberation time, T , and the directivity coefficient, γ . In the subsequent calculation, omnidirectional sound sources are assumed to have a directivity coefficient that equals to one.

In the next step, the impulse response is calculated at a receiver position at the critical distance. The overall energy of the impulse response, E_T , is the sum of the direct sound energy, E_D , the early reflection energy, E_E , as well as the late reverberant energy, E_L , as follows,

$$E_T = E_D + E_E + E_L. \quad (7)$$

At the critical distance, the following condition has to be met for an omnidirectional source and receiver pair,

$$E_D = E_E + E_L. \quad (8)$$

Consequently, the energy of the late reverberation has to be adjusted to

$$E_L = E_D - E_E, \quad (9)$$

$$\sum_{t=0}^{2RT} p_L^2 = \sum_{t=0}^{2RT} p_D^2 - \sum_{t=0}^{2RT} p_E^2, \quad (10)$$

with the sound pressure, p , which is, of course, proportional to the digital signal amplitude.

In the second method, the exponentially-decaying amplitude of the reverberation tail is fitted to the exponentially-decaying amplitudes of the reflection pattern. For this purpose, both signals are logarithmized so that the decaying impulse response can be fitted by a linear-regression curve. The amplitudes of the decaying slope are then matched and a cross-fade method is used to blend out the early reflections while gradually blending in the late reverberation.

2.3 The Forest Walk-Through

Coming back to the forest simulation example, the environment is depicted in Fig. 2b. The red dot shows the sound source, which is located at the coordinate 40 m/7 m (x/y coordinates in meters). In cases where two circles overlapped, one of the circles was removed since two tree stems cannot occupy the same space. The absorption coefficient was set to 5% based on measurement results for tree barks (Reethof et al. 1977). Diffuse reverberation that results from leaves and other objects was added. The reverberation time was adjusted to 1.6 s and the reverberation ratio, adjusted to the interaural coherence, was 0.4 at a source-to-receiver distance of 40 m. Both values were chosen based on forest-acoustics measurements by Sakai et al. (1998). The rays are depicted through dashed lines that become lighter with increasing order. Three walk pathways were computed at different y-coordinates that were held constant for each condition, that is, 17 m, 27 m, and 47 m, labeled as 10-m, 20-m and

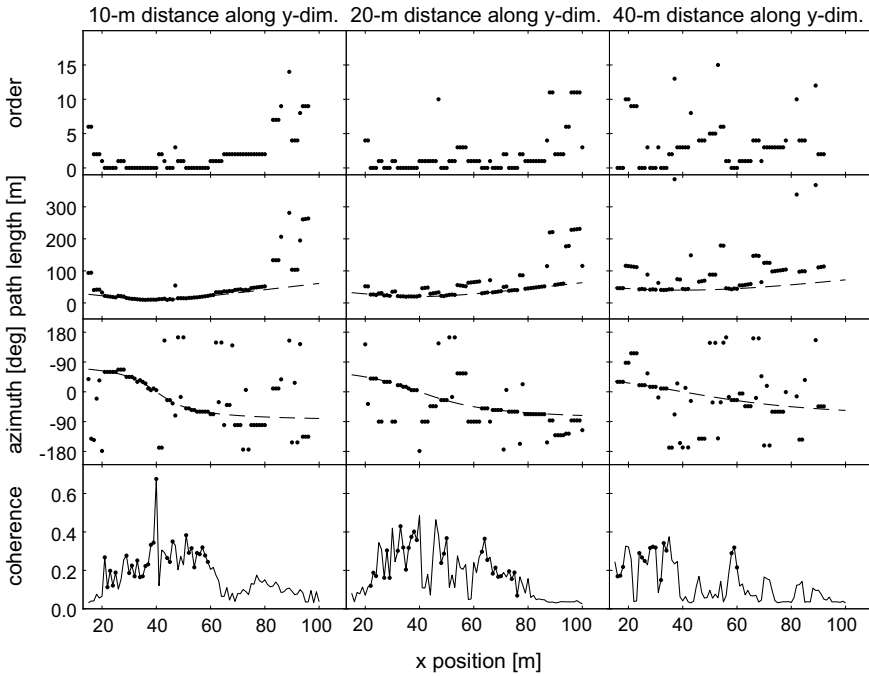


Fig. 4 Results of cue analysis of a simulated forest walk-through. The **left panels** show the results for the pathway that is 10m behind the listener in the y-direction. The **center panels** show the results for the 20-m condition, and the **right panels** for the 40-m condition. For each condition, the order of the first arriving reflection is shown in the top graph as a function of the x-position. An order of zero indicates that the direct signal passes to the listener position and is not obstructed by the tree trunks. The **panels 2nd from the top** show the actual path lengths of the first wavefront from the source to the receiver—indicated by the **dots**. The **dashed line** shows the direct distance between the sound and the listener. The **panels 3rd from top** show the angles of incidence of the first-arriving wavefront indicated by the **dots**. The **dashed lines** show the actual azimuths between the sound source and the listener. The **bottom graphs** depict the coherence indicated by the **solid lines**. All coherence values that correspond to cases where the direct signal was not obstructed are emphasized by additional **dots**

40-m conditions—referring to the distance in the y-dimension between source and receiver. Each walkway covers the distance from 15 to 100m along the x-coordinate. Binaural impulse responses were computed in 1 m increments along the x-axis and then analyzed.

The analysis results are shown in Fig. 4. The three columns show the results for the different distances along the y-coordinate. The top row shows the order of the first arriving wave. For the 10-m condition, the unobstructed direct signal (0th order) arrives at the listener position in 37% of the cases—see top-left graph of Fig. 4. In those cases, where the receiver is further away from the sound source based on the x-axis position (x-position > 60 m), the direct line of sight is obstructed in all cases and the first ray that reaches the receiver is typically on the order of two or higher.

For $x > 80$ m, the effective path length becomes much greater than in the other cases with values of 100 m and beyond—see graph second from the top in the left panel of Fig. 4. This also greatly affects the azimuth of the first arriving wavefront, which is not necessarily the direct signal—see graph second from the bottom in the left panel of Fig. 4. With a few exceptions, the first wave front arrives from the azimuth direction of the sound source or an angle close by if the receiver is located at an x -position between 20 and 60 m. Outside this range, the azimuth values differ greatly from the actual sound-source angle.

Next, the interaural coherence is investigated. The interaural coherence estimates how similar the left and the right ear signals are in the time domain after they have been amplitude and time aligned. It is a measure of how reverberant the sound field is at the listener position—knowing that the presence of reverberation decorrelates both signals, thus making them more dissimilar. The interaural coherence can be calculated as the absolute maximum of the normalized cross-correlation function, which is defined as

$$R_{l,r}(n, m) = \frac{\sum_{n=n_0}^N x_l(n - m) \cdot x_r(n)}{\sqrt{\sum_{n=n_0}^N x_l^2(n - m) \cdot \sum_{n=n_0}^N x_r^2(n)}}, \quad (11)$$

with the time n , the internal delay, m , the left input signal, x_l , and the right input signal, x_r .

The interaural coherence for the 10-m condition is shown in the bottom-left graph of Fig. 4, solid line. It is noteworthy that the interaural coherence becomes noticeably larger with the absolute distance from the source. Therefore the interaural coherence is generally smaller in the 20-m–60-m x -position range than for the outside positions. All values that correspond to cases that include the direct sound are emphasized through dots.

The 20-m condition is shown in the center column of Fig. 4. The relative number of x -positions, where the direct signal is not obstructed on the pathway to the listener position is slightly lower than in the 10-m condition—33% versus 37%. The average coherence, 0.17, is the same as found for the 10-m condition. Also, the coherence is noticeably higher for most cases, where the direct signal reaches the listener's ears—as indicated by the dots.

The 40-m condition is shown in the right column of Fig. 4. Here, the direct path is often obstructed and the direct signal reaches the listener only in 17% of the cases. Consequently, only a few azimuth values indicate the correct sound-source position—see graph second from the bottom in the right panel of Fig. 4. The coherence values are lower than for the other two conditions with an average of 0.13. However, also in this case, the relative coherence values are higher when the direct signal is not obstructed on its way to the listener.

In general, it can be concluded that in a dense forest environment the direct signal is often obstructed, making it both acoustically and visually challenging to localize

salient objects. By moving through the environment, the receiver can find locations where the sound source arrives directly without obstructions. These positions are usually characterized by coherence values that are higher than those found for obstructed sound sources. The forest scenario is a good example of where hearing and vision can work together to locate sound sources quickly because the sound source becomes visible once it is no longer obstructed by objects.

3 Understanding the Fundamental Sound of Caves to Concert Halls Using a Precedence-Effect Model

The next section investigates how binaural models can be used to extract room-acoustic features from a running signal and compare the results to the first-known type of concert venue—the cave. At the core of the human ability to extract information from sound sources in reverberant spaces are auditory mechanisms related to the *precedence effect* (Blauert 1997; Litovsky et al. 1999). The precedence effect, formerly also called the *law of the first wave front*, describes the ability of the auditory system to suppress information about secondary sound sources that are reflected from walls and other objects. This enables the auditory system to localize the actual position of a sound source by making the localization cues pertinent to the direct-signal component available. This is a non-trivial task for the auditory system since the direct signal and the reflected signal parts overlap in time and frequency. The primary cues to localize a sound source are *Interaural (arrival) Time Differences* (ITDs) and *Interaural Level Differences* (ILDs). ITDs occur because the path lengths between a sound source and both ears differ depending on the incoming azimuth angle. The cross-correlation algorithm, (11), is an adequate algorithm to simulate the processes in the auditory system when extracting ITD cues. The lateral position of the cross-correlation peak as a function of the internal delay, m , is used to determine the ITD. ILDs occur because of shadowing effects of the head toward the contralateral ear. For more details on ILDs, see, for instance, Breebaart et al. (2001), Braasch (2003, 2005).

The third type of spatial cues are called “*monaural cues*”. Monaural cues are direction-dependent, pinna-induced spectral modifications that require only one ear for analysis—Blauert (1969/1970, 1997), Zakarauskas and Cynader (1993). These cues are especially important for judging the elevation and front/back orientation of a sound source. Yet, it is shown in this chapter that these dimensions can also be handled by ITD-based algorithms if head movements are considered—compare Pastore et al. (2020, this volume). The focus of the chapter will, however, continue to focus on sound-source localization and information extraction in reverberant environments. Further, regards survival, the auditory system’s ability to segregate sound sources is relevant as well—for details see, for instance, Bodden (1993), Roman et al. (2003), Roman et al. (2006), Deshpande and Braasch (2017), Mi et al. (2017).

When conducting room analyses, it should be kept in mind that modern concert halls have not been around until very recently in the scheme of human history. It will now be discussed how the auditory system can extract room-acoustic features by using a system that does not have at its disposal auditory experience of millions of years to adapt to rectangular-shaped rooms.

The first-known musical instrument at all is a 40,000-year-old rim flute made from a vulture bone. Yet, in the context of the current paper is not so much the instrument itself that is important but rather the cave where it was found. This is the *Hohle-Fels* cave near Schelklingen, Germany—Conard et al. (2009). It is hard to imagine that this instrument has not been played in the cave. In 2016, the current author had the opportunity to visit the cave and to record some impulse-responses in it. This allowed him to estimate the cave's mid-frequency reverberation time, which came out as about 2 s—Braasch (2019). It is remarkable that the reverberation time of the *Hohle-Fels* cave is in the range of modern classical concert halls. For example, the Haydn-Saal in Eisenstadt, as was shown in Fig. 1, also has a reverberation time of about 2 s in the mid-frequency range when the hall is occupied—Meyer (1978), p.147. However, what is important in the context of this paper, is the following. Despite the similarities in reverberation times of the *Hohle-Fels* cave and a typical concert hall, there is a fundamental acoustic difference between them. Concert halls are typically rectangularly shaped or have at least large plane surfaces, while the surface of a cave is very irregular. The latter leads to a very diffuse echogram while the concert hall has a few very distinct reflections. While reverberation chambers for technical acoustic measurements are often kept diffuse, there are very few music facilities that build on this diffuseness notion. The most distinct two in the world are probably Studio C at Blackbird Studios in Nashville, TN, (Bonzai 2018), and the Studios 1 and 2 at Rensselaer Polytechnic Institute's Experimental Media and Performing Arts Center (EMPAC) in Troy, NY. Studio C was conceived and designed by George Massenburg, and its walls are treated with 40 ton of long wood beams similar to the absorptive wedges in anechoic chambers but with an irregular pattern and being sound-reflective. While the sound of the studio is reverberant, this unique design avoids spectral colorations imposed by comb filtering. Concurrently, a lot of spatial properties that would commonly originate from the pattern of specular reflections are not present in this studio. An anecdote illustrates the acoustic features of this space. According to George Massenburg, a session was booked with a blind pianist. When the musician entered the studio, he walked around the music stand and the piano with the help of his cane but then, without the usual direct reflections of planar surfaces, walked straight into a wall that he did not perceive to be there. Studio 1 and 2 at EMPAC were conceived by EMPAC director Johannes Goebel with the diffuse acoustics of a forest opening in mind.

3.1 Precedence-Effect Model

Estimating the ITD of the Direct Signal

The precedence-effect model as reported here, has the task of analyzing the reverberant conditions. It is on the *Binaurally Integrated Cross-correlation Auto-correlation Mechanism (BICAM)*—Braasch (2016). Modifications were made to the original algorithm to calculate more accurate binaural-activity maps. Unlike traditional precedence-effect models that suppress the energy or spatial information of early reflections, the BICAM algorithm separates the auditory cues for the direct signal and from the early reflections but does not remove or suppress the latter. This is important for this section because the aural quality of the room that the sound source is presented in can thus be evaluated. The model separates auditory features for the direct signals and early reflections from a running signal using a dual-layer spatiotemporal filter. Figure 5 shows the architecture of the model. The incoming signal ascends from bottom to top. The model separates the incoming binaural signal into auditory bands at the initial stage—as shown in the bottom row of boxes. Then, the model performs a set of auto-/cross-correlation analyzes within all auditory bands as depicted in boxes, labeled “AC” and “CC”, that are shown in the 2nd row from the bottom. During this process, the following autocorrelation/crosscorrelation sequences are calculated from the left and right ear signals, x and y —depicted as Steps 1 and 2 in Fig. 6,

$$R_{xx}(m) = E[x_{n+m}x_n^*] \quad (12)$$

$$R_{yx}(m) = E[y_{n+m}x_n^*] \quad (14)$$

$$R_{xy}(m) = E[x_{n+m}y_n^*] \quad (13)$$

$$R_{yy}(m) = E[y_{n+m}y_n^*], \quad (15)$$

with the cross-correlation sequence, R , the expected-value operator, $E\{\dots\}$. The regular, non-normalized cross correlation is defined as follows:

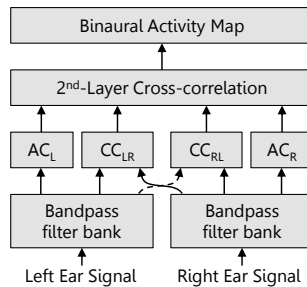


Fig. 5 System architecture and signal flow of the BICAM model—(AC)...autocorrelation, (CC)...cross correlation

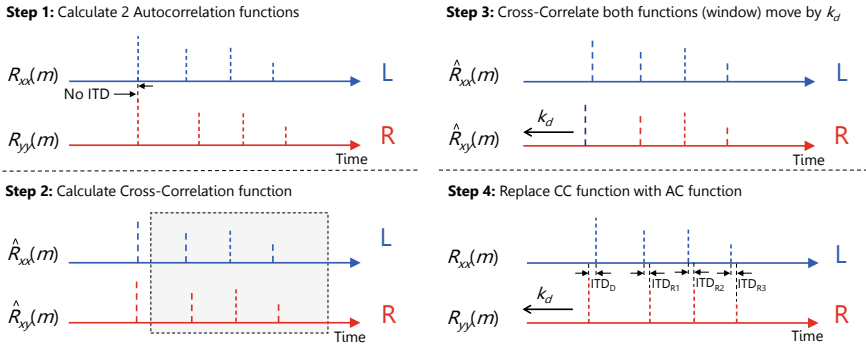


Fig. 6 Autocorrelation/cross-correlation procedures that are performed using the BICAM architecture to estimate the binaural room impulse response. The variable $R_{x,x}$ is the autocorrelation for the left channel, $R_{y,y}$ denotes the same for the right channel. $R_{x,y}$ represents the cross-correlation function between the left and the right channels. The hats over the variable R indicates that only the right side is considered. The **gray window** in Step 2, is used to compare the left-ear channel, L , to the right-ear channel, R

$$R_{i,j}(n, m) = \sum_{n=n_0}^N i(n - m) \cdot j(n), \tag{16}$$

with, the time, n , the internal delay, m , and the input signals i, j . The variable n_0 is the start time of the analysis window and N the end time. The left and right input signals are assigned to the variables i and j . For the case $i = j$, R denotes the autocorrelation. For the BICAM model, the range of the internal delays, $-M$ to M , needs to exceed the duration of the reflection pattern of interest. Otherwise, the impulse response is not shown in its entire duration. Alternatively, $\pm M$ can also be set to just show the early part of an impulse response. The variable n typically ranges from the beginning of the signal, $n = 0$, to the end of the signal, N . The calculation can be performed as a running analysis over shorter, overlapping time segments.

In the next step, which is typically not found in traditional cross-correlation models (Sayers and Cherry 1957; Blauert and Cobben 1978; Stern and Colburn 1978), a cross-correlation algorithm is performed on top of the combined autocorrelation/cross-correlation algorithm as shown in the second top box in Fig. 5 and also in Step 3 of Fig. 6. The goal of this procedure was to develop a method that incorporates the causality of the direct sound and its reflections, which is not provided by conventional cross-correlation models. Using the second-layer cross-correlation analysis over the autocorrelation signal (e.g., $R_{x,x}$) in one-channel and the cross-correlation signal (e.g., $R_{x,y}$) in the second channel, the spatial information in the direct signal and in the individual reflections can be segregated.

A key to the function of the model is a comparison of the right side peaks of both functions (autocorrelation function and cross-correlation function) as shown in the gray box in Step 2 of Fig. 6. These side peaks are correlated to each other by windowing out the direct peaks and the left side of the (auto-)correlation func-

tions. The temporal offset between both main peaks can be obtained by aligning the side peaks in time to determine the interaural time difference (ITD) of the direct sound. The alignment of the side peaks is accomplished by cross-correlating the two autocorrelation/cross-correlation functions—(12) and (15)—over the segments of both functions that contain the side peaks for positive internal delay values, m , (gray areas in Step 2 of Fig. 5). It is important to zero out all remaining segments so that the main peaks and the side peaks for negative m values cannot affect the alignment of the positive side peaks. Mathematically, this operation can be stated as

$$\hat{R}_{ij} = R_{ij} | \forall m > w \wedge \hat{R}_{ij} \stackrel{!}{=} 0 | \forall -M \leq m \leq w. \quad (17)$$

In the next step, the variables, i and j , are substituted with the left and right ear signals, x and y , to compute the following four functions, \hat{R}_{xx} , \hat{R}_{xy} , \hat{R}_{yx} , and \hat{R}_{yy} . The variable w is the length of the window to remove the main peak. The method works if the cross terms (correlations between the reflections) are within certain limits.

Using these functions, the 2nd-layer cross-correlation is calculated. The ITD for the direct signal, $k_{\bar{d}}$, can then be computed from the product of the 2nd-layer cross-correlation terms—see Step 3 in Fig. 5:

$$k_{\bar{d}} = \max_m \arg \left\{ \sqrt{|R_{\hat{R}_{xy}\hat{R}_{xx}} \cdot R_{\hat{R}_{yy}\hat{R}_{yx}}|} \right\}. \quad (18)$$

The solution for $k_{\bar{d}}$ represents the lateral position of the direct signal. In the next step, this solution is used to further expand the algorithm to derive a binaural-activity map that also contains information about the locations and delays of individual early reflections—see top box in Fig. 5.

Binaural-Activity-Map Calculation

A binaural activity map is a three-dimensional plot of a binaural room impulse response that depicts the temporal course of the reflections on the x-axis, the spatial positions of the reflections on the y-axis and the amplitude of the reflections on the z-axis—see Braasch (2005) for more information. In order to create the binaural activity map, the ITD of the direct signal, $k_{\bar{d}}$, is used to shift one of the two autocorrelation functions, R_{xx} or R_{yy} . The latter two functions are, in some form, a representation the early reflection patterns for the left and right channels—see Step 4 in Fig. 5. The respective equations are

$$\check{R}_{xx}(m) = R_{xx}(m), \quad (19)$$

$$\check{R}_{yy}(m) = R_{yy}(m - k_{\bar{d}}). \quad (20)$$

A series of cross-correlation functions is calculated over moving segments of the time aligned autocorrelation functions, \check{R}_{xx} and \check{R}_{yy} , for positive time values in order to estimate the delays, ITDs and relative amplitudes of the reflections.

3.2 Acoustical Analysis

In order to demonstrate the effects of the two opposite sound environments, two idealized environments were created, one with mostly diffuse reflections and the other one with mostly specular reflections, using the ray-tracing model that was introduced in Sect. 2.2. In the first case, simulating the cave, the impulse response for the diffuse reverberation of the *Hohle-Fels cave* was simulated using decaying Gaussian-noise burst roughly matching the RT of the cave, namely, 2 s, and an initial time-delay gap of 12 ms. The direct sound source was simulated using a delta peak convolved with an HRTF pair corresponding to 0° azimuth and 0° elevation. All HRTF catalogs used for this chapter have been measured at the Institute of Communication Acoustics of the Ruhr-University Bochum, Germany—for details see Braasch and Hartung (2002). The direct-to-reverberant-energy ratio between the direct signal and the late reverberation was set to 3 dB. An anechoic male voice sample of 12 s duration was used as the sound signal for all examples in this section. For the simulation, the auto- and cross-correlation terms, \hat{R}_{xx} , \hat{R}_{xy} , \hat{R}_{yx} , and \hat{R}_{yy} of the BICAM algorithm were employed. The values for Eq. (17) are calculated in separate auditory bands using the same gammatone-filter bank (Patterson et al. 1995) with 15 auditory bands from 100 to 1600 Hz. The beginning of the window w in (17) was set to 100 samples (2268 μ s). The length of the window equaled to 40 ms. The ITD, $k_{\bar{d}}$, for the direct signal was then estimated from the frequency-superposed 2nd-layer cross-correlation functions according to (18).

Figure 7a shows the binaural impulse response extracted from the running signal using the BICAM model. One can clearly see that both the left and the right channels, shown as blue and red lines, depict the direct sound but not the exponentially decaying reverberation tail. Only some residual noise that results from the autocorrelation process can be found due to the limited duration of the source signal. The binaural-activity map that was computed using the estimated binaural impulse response is shown in Fig. 8a. Based on the discussed features of the impulse response, it comes as no surprise that the binaural-activity map only shows a single peak for the direct signal but no trace of the reflections. With the exceptions of a few artifacts at the late end of the binaural activity map, the outcome is very similar to the binaural-activity map computed for an anechoic environment but an identical direct sound source—see Fig. 7b and Fig. 8b. The results support the StudioC anecdote and are also in line with a study by Teret et al. (2017) that demonstrates that listeners have no temporal representation of a Gaussian reverberation tail independent of the sound stimuli that are convolved with that reverberation tail.

For the concert-hall example, the impulse response was composed of the same direct signal and its reverberant tail as has already been used to simulate the cave. In addition, four specular reflections were simulated at the following locations and delays,

Azimuth	Delay	Reflection coefficient
-45°	16 ms	0.7
+45°	19 ms	0.7
-60°	22 ms	0.5
+60°	25 ms	0.4

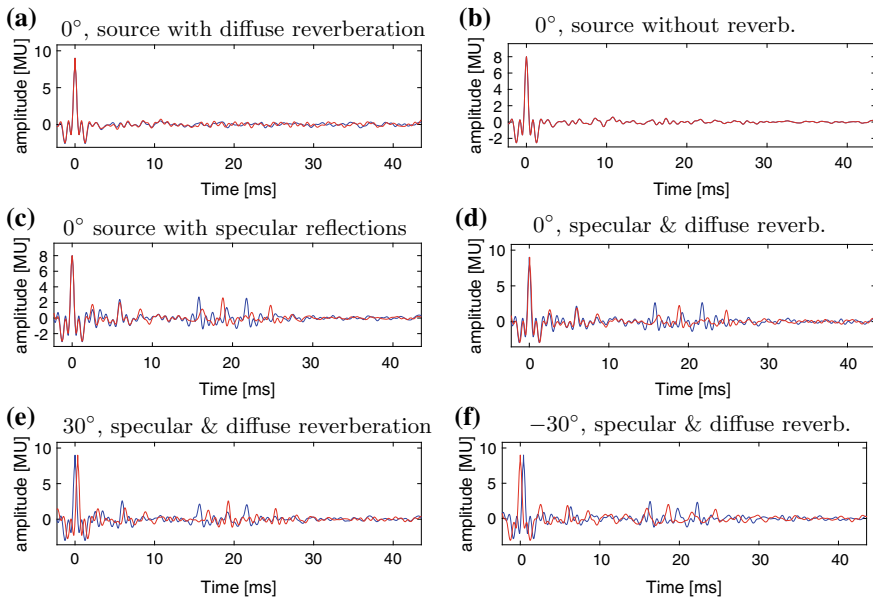


Fig. 7 Binaural room impulse responses, estimated from running signals

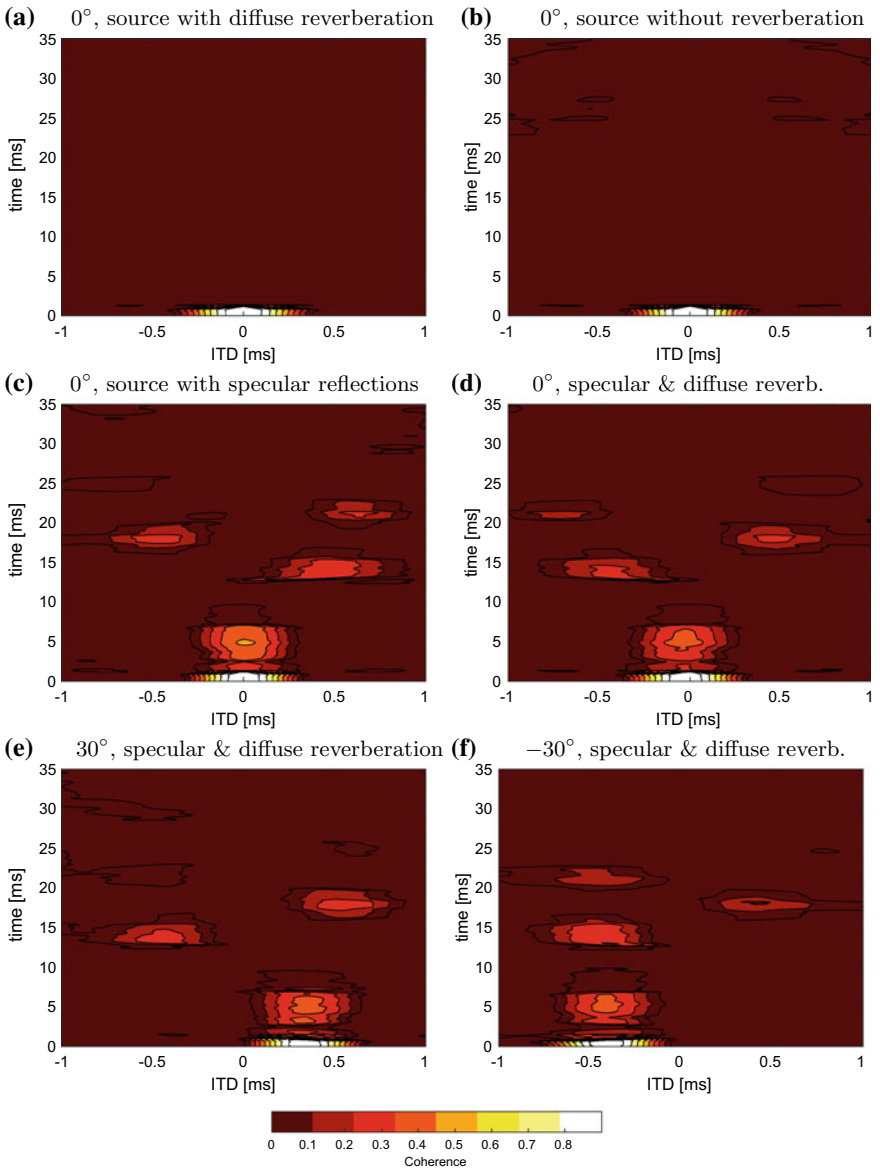


Fig. 8 Binaural-activity-map results of BICAM-model analyses for different conditions, including diffuse and specular reflections as indicated in the individual graphs above

The extracted impulse responses are shown in Fig. 7c,—impulse response direct sound and specular reflections only—and in Fig. 7d,—impulse response with specular reflections and diffuse reverberation tail. The corresponding binaural-activity maps are shown in Fig. 8c, d. In both cases, the map correctly identifies the lateral positions and the delays of the direct sound and of most of the early reflections. In order to demonstrate the ability of the model to provide an independent but joint analysis of the direct sound and the early reflections, only the lateral position of the direct sound source was moved while the lateral positions and delays of the early reflections were maintained. Figure 8e shows the same binaural-activity map as Fig. 8d but for a direct signal that has been moved laterally to $+30^\circ$. The result shows that the binaural-activity map indicates the new lateral position correctly while maintaining, in principle, the positions of the side peaks that indicate the delays and lateral positions of the reflections. The positions of the side peaks are also maintained when the direct-sound source is moved to -30° —see Fig. 8f. Figure 7e, f depict the binaural room impulse responses that were extracted from the running signal for the two conditions with a lateralized direct-sound source. In comparison with the laterally centered direct-sound source, shown in Fig. 8d, one can see that the later parts of the binaural room impulse responses are very similar while the onset delays between left and right channels, shown in blue and red, are clearly visible.

Before concluding this section, the fundamental differences of the BICAM algorithm when processing specular reflections and diffuse reflections should be discussed. In this context it is worth noting that the binaural-activity map for the condition with laterally centered direct signal and early specular reflections, as shown in Fig. 8c, does not change much when a late, diffuse reverberation tail is added—see Fig. 8d. Both maps are very similar indeed despite the fact that the stimulus of Fig. 8d contains a diffuse reverberation tail in addition to the early specular reflections. The similarity in both maps re-emphasizes that the BICAM method is “blind” toward diffuse reverberation tails because these do not produce a distinct autocorrelation map. Obviously, human listeners are aware of the presence of a late reverberant field, otherwise acoustical designs like the Blackbird’s StudioC and the EMPAC’s Studios 1 and 2 would not have a meaningful purpose. It has to be kept in mind, though, that the proposed BICAM model is a localization model. Other types of psychoacoustic models, such as detection models, are needed to extract further room-acoustic features. Further standard methods estimate, for instance, interaural coherence and/or extract features of the (exponentially) decaying room impulse response from transients in the source signals, in particular, from impulses and abrupt stops.

While the interaural-coherence method is usually calculated from a measured room impulse response, it can also be calculated from a running signal. This was done for the forest walk-through—see (11). The drawback using the latter method is that the type of source signal will influence the interaural coherence, and the outcome is no longer solely based on the room parameters. However, also in real life, the perceived reverberance is highly influenced by the source signal employed—Teret et al. (2017). A further method is to estimate the reverberation time from the exponential-decay rate—see, for instance, Huang et al. (1999).

Among the three described methods, the binaural-activity-map analysis is the only method that allows for the extraction of information about the location (angle and distance) of reflective surfaces, for instance, of walls. Neither the interaural-coherence method nor the exponential-decay method provides these cues. Without them, the listener will not receive unambiguous information about the size of a room and the location of walls and of further sound-reflecting or obscuring obstacles. This is the reason why the blind pianist walked right into a wall in Blackbird's Studio C—the absence of salient cues.

4 Simulating an Office Walk-Through Using a Binaural Model Capable of Utilizing Head Movements

An acoustic walk-through a building is in many ways a modern-society version of the forest walk-through. Also in this case, the direct sight to an object can be obstructed and then one has to rely on the acoustic sense. However, obstructing objects have very different acoustic qualities. While the forest is a leaky reverberation chamber with diffusive character, office suites, and other small rooms are characterized by specular reflections that arrive shortly after the direct sound. From an evolutionary perspective, where anatomical changes occur over a span of several million years, rectangular caverns with flat walls have only been introduced recently during our early civilization and similar acoustic objects do not appear in nature. It is therefore important to understand how the auditory system is able to adapt to built rooms given that it has not specifically evolved to deal with such environments.

4.1 Head-Movement Algorithm

In order to simulate the office walk-through, an existing head-movement algorithm (Braasch et al. 2013) is added to the binaural model, such that the model can resolve back/front confusions and analyze the auditory scenes adequately. The model builds on a theory proposed by Wallach (1939).

Auditory Periphery

The model takes a step back from the elaborate BICAM mechanism and uses a traditional interaural-cross-correlation method as introduced by Sayers and Cherry (1957) to estimate ITDs—see (17). The basic model structure, shown in Fig. 9, is similar to the one proposed by Braasch (2002). The inputs signals are filtered with HRTFs from desired directions. Basilar-membrane and hair-cell behavior are simulated using a gammatone-filter bank with 36 bands and a simple half-wave rectifier at a sampling frequency of 48 kHz, as described by Patterson et al. (1995).

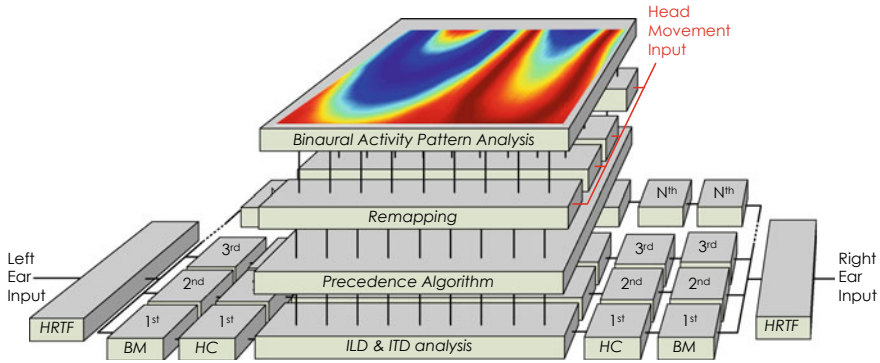


Fig. 9 General model structure of the binaural-localization model utilizing head rotations. **HRTF**...External-ear simulation/HRTF filtering. **BM**...Basilar membrane/bandpass filtering. **HC**...Hair cell/halfwave rectification **ITD & ILD analysis**...Interaural time difference-cue (ITD) extraction/interaural cross-correlation and interaural level-difference cue (ILD) analysis with EI-cells, precedence effect algorithm, remapping to azimuths with head-rotation compensation, binaural-activity-map analysis to the estimation the sound-source positions

Cross Correlation

After the half-wave rectification, the normalized interaural cross-correlation (11) is computed for each frequency band over a short time segment. Only the Frequency Bands 1 to 16 (23–1559 Hz) are analyzed, reflecting the inability of the human auditory system to resolve the temporal fine structure at high frequencies, as well as the fact that at low frequencies the interaural time differences in the fine structure are the dominant cues—provided that they are available at all (Wightman and Kistler 1992).

Remapping and Decision Device

Next, the cross-correlation functions will be remapped from interaural time differences to azimuth positions. This is important for the model to be able to predict the spatial position of the auditory event. In addition, this procedure helps to align the estimates for the individual frequency bands as one cannot expect that the interaural time differences are constant across frequency for a given angle of sound-source incidence. An HRTF catalog is analyzed to convert the cross-correlation function’s x-axis from interaural time differences to the azimuth. The HRTF catalog was measured at a resolution of 15° in the horizontal plane and then interpolated to 1° resolution using the spherical-spline method—see Hartung et al. (1999). After filtering the HRTFs with the gammatone-filter bank, the ITDs for each frequency band and angle are estimated using the interaural-cross-correlation (ICC) algorithm of (16). This frequency-dependent relationship between ITDs and azimuths are used to remap the output of the cross-correlation stage (ICC curves) from a basis of ITDs $m(\alpha, f_i)$, to

a basis of azimuth angles in every frequency band as follows:

$$m(\alpha, f_i) = g(\text{HRTF}_l, \text{HRTF}_r, f_i) \quad (21)$$

$$= g(\alpha, f_i), \quad (22)$$

with azimuth, α , elevation, $\delta = 0^\circ$, distance, $r = 2$ m, $\text{HRTF}_{l/r} = \text{HRTF}_{l/r}(\alpha, \delta, r)$, center frequency of bandpass filter, f_i .

Next, the ICC curves, $(R_{x,r}(m, f_i))$, are remapped to a basis of azimuths using a simple for-loop in Matlab using a step size of 1° :

```
for alpha=1:1:360
    R_rm(alpha, freq)=R(g(alpha, freq), freq);
end
```

Here, $R(m, \text{freq})$ is the original, frequency dependent, interaural-cross-correlation function with the internal delay, m . The function $g(\alpha, \text{freq})$ provides the measured m -value for each azimuth and frequency. Inserting this function as input, m , to R transforms the R -function into a function of the azimuth, using the specific Matlab syntax.

In the decision device, the average of the remapped ICC functions, $R_{\text{rm}}(\alpha, \text{freq})$, over the frequency bands 1–16 is calculated and divided by the number of frequency bands. The model estimates the sound sources at the positions of the local peaks of the averaged ICC function.

Figure 10 shows an example of a sound source in the horizontal plane with an azimuth of 30° for the eighth frequency band. The top-left graph shows the original ICC curve obtained using (16) as a function of ITD. The graph is rotated by 90° with the ICC on the x-axis and ITD on the y-axis to demonstrate the remapping procedure. The curve has only one peak at an ITD of 0.45 ms. The top-right graph depicts the relationship between ITD and azimuth for this frequency band. As mentioned previously, the data were obtained by analyzing HRTFs from a human subject. Now, this curve will be used to project every data point of the ICC-versus-ITD function to an ICC-versus-azimuth function, as shown for a few data points using the straight dotted and dashed-dotted lines. The bottom panel shows the remapped ICC function, which now contains two peaks, that is, one for the frontal hemisphere and one for the rear hemisphere. The two peaks fall together with the points where the cone-of-confusion hyperbolas intersect the horizontal plane for the ITD value of the maximum peak that is shown in the top-left panel.

Integrating Head Rotation

In the following, it is assumed that the head rotates to the left while analyzing an incoming sound source from the front. Related to the head, the sound source will move toward the right. However, in the case that sound source was in the rear, the

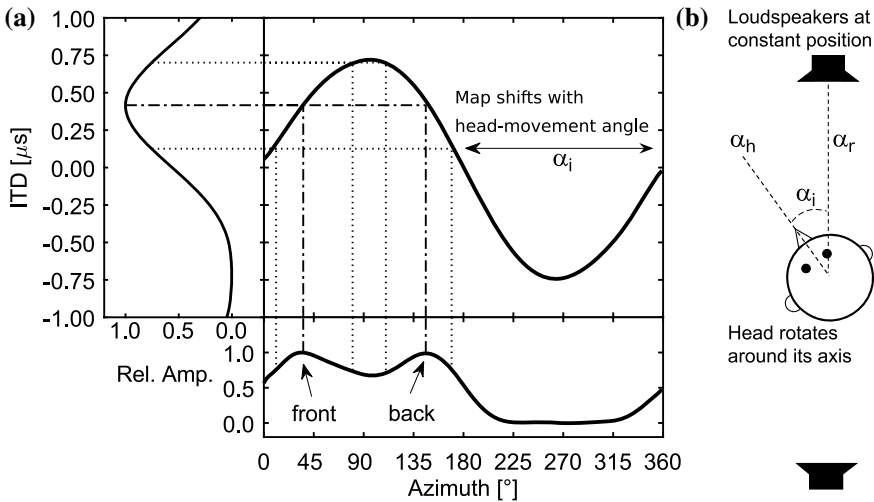


Fig. 10 (Left) Remapping of the cross-correlation function from ITD to azimuth, shown for the frequency band 8, centered at 434 Hz. The signal was presented at 30° azimuth, 0° elevation. **(Right)** Sketch to illustrate the front/back confusion problem. If an ongoing sound source is located in front of the listener who turns her head left, the sound source will move to the right from the perspective of the listener’s head. But if the sound source is located in the back, the sound source appears to move to the left for the same head rotation. The variable, α_r , denotes the azimuth in the room-related coordinate system, pointing here at 0°. The variable α_h is the azimuth in the head-related coordinate system, also pointing at 0° but for this coordinate system. The third angle, α_m , is the head-rotation angle, which indicates by how much the head is turned from the reference head orientation that coincides with the room-related-coordinate system

sound source had moved to the left. This phenomenon will now be used to distinguish between both options, that is, frontal and rear position. For this purpose, a different coordinate system is introduced, namely, the room-related coordinate system. The fact that human listeners maintain a good sense of the coordinates of a room as they move through it, motivates this approach. If a stationary head position is considered, the head-related coordinate system is fully sufficient. However, if the head rotates or moves, the description of stationary sound-source positions can become challenging because every sound source starts to move with alterations of the head position. An easy way to introduce the room-related coordinate system is to define a reference position and reference orientation of the human head, and then determine that the room-related coordinate system coincides with the head-related coordinate system for the chosen reference position—compare Pastore et al. (2020, this volume), for details on this topic, involving multimodal cues.

Consequently, the room- and head-related coordinate systems are identical if the head does not move. In this investigation, only head rotations within the horizontal plane are considered, and for this case, the difference between the head-related coordinate system and the room-related coordinate system can be expressed through the head-rotation angle α_i that converts the room-related azimuth α_r to the head-related

azimuth α_h —see Fig. 10, right graph. That is,

$$\alpha_r = \alpha_i + \alpha_h. \quad (23)$$

Given restricted head movement, the origin of both coordinate systems and the elevation angles are always identical. While the sound-source position changes relative to the head with head rotation, a static sound source will maintain its position in the room-related coordinate system. Using this approach, another coordinate transformation of the ICC function is executed in the model, namely, a transformation from of head-related to room-related azimuth. This can be accomplished by rotating the remapping function when the head is moving by $-\alpha_i$ to compensate for the head rotation.

If a physical binaural manikin were used—with a motorized head in connection with the binaural model—the HRTF would be automatically adjusted with the rotation of the manikin’s head. In the model discussed here, where the manikin or human head is simulated by means of HRTFs, the HRTFs have to be adjusted virtually. Also, at every moment in time the HRTFs have to correspond to the sound-source angle relative to the current head position. This can be achieved with the help of a running window function, where the sound source is convolved with the current HRTF pair. A Hanning window of 10 ms duration and a step size of 5 ms is used here for this purpose. The smooth edges of this window will cross-fade the signal allowing a smooth transition during the exchange of HRTFs. For each time segment, the model processes the following sequence:

1. First, it updates the current head-rotation angle, α_i
2. Then it calculates the current head-related azimuth angle, α_h , for each sound source located at its room-related azimuth, α_m
3. Next, the model selects the HRTF pair that correspond closest to α_h
4. Afterwards, it computes the normalized ICC, $R_{l,r}$, for each frequency as a function of the ITD
5. It converts the ICC function to a function of head-related azimuth, α_h , using the remapping function shown in Fig. 10.
6. Next, the model circular-shifts the remapping function based on the head-rotation angle by $-\alpha_i$ to transform the ICC curve into the room-related coordinate system
7. Then, it computes the mean ICC output across all frequency bands
8. It averages the ICC outputs over time
9. It estimates the position of the auditory event to be at the azimuth where the ICC peak has its maximum

The first example is based on a bandpass-filtered white-noise signal with a duration of 70 ms. The signal is positioned at -45° azimuth in the room-related coordinate system. At the beginning of the stimulus presentation, the head is oriented toward the front, $\alpha_h = 0^\circ$, and then rotates with constant angular velocity to the left until it reaches an angle of 30° at the time that stimulus is turned off. The ICC functions are integrated over the whole stimulus duration. Figure 11 shows the result of the simulation. The initial ICC-versus- α_r function the output of Step 6 for $\alpha_m = 0^\circ$ is

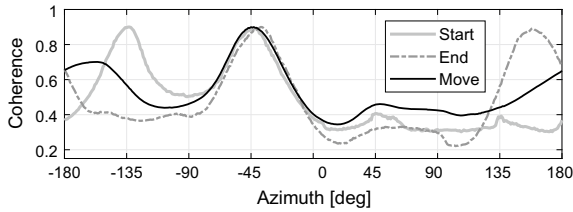


Fig. 11 Interaural-cross-correlation pattern for a sound source at -45° which is presented during a head rotation from $\alpha_m = 0^\circ$ to 45° . The **dashed line** shows the ICC curve for the initial time window, the **solid gray curve** for the last segment when the head is fully turned. Note that the ICC pattern was shifted in the opposite direction of the head rotation to maintain the true peak position at $\alpha_r = -45^\circ$. The **black curve** shows the time-averaged ICC curve for which the main ICC peak remains and the secondary ICC partly dissolves

depicted by the solid, light gray curve. Here clearly two peaks can be observed, one at $\alpha_{r=h} = -45^\circ$ and another one at $\alpha_{r=h} = -135^\circ$. At the end of the stimulus presentation shown as the dashed, dark gray curve $t = 70$ ms, $\alpha_m = 45^\circ$ —, only the position of the rear peak is preserved. This peak indicates the “true”, that is, the physical sound-source location, $\alpha_{r \neq h} = -45^\circ$, because the head rotation was compensated for by rotating the remapping function in opposite direction of the head movement.

However, in the case of a front peak, that is, the front/back confused position, the peak position was counter-compensated for and it rotates twice the value of the head-rotation angle, $\alpha_m = 30^\circ$. The new peak location is shifted by -60° to a new value of 165° . The time-averaged curve (the solid black line which shows the output of Step 7) demonstrates the model’s ability to robustly discriminate between front and rear angles. The secondary peak, the one representing the solution for a frontal sound source, is now smeared out across the azimuth because of the head rotation. Further, its peak height is reduced from 0.9 to 0.7 making it easy to discriminate between front and rear.

4.2 Analysis of an Office Walk-Through

In the next example, it is investigated how the combined head-movement and BICAM localization models can be applied to a real-world scenario, for example, to sound localization in an office suite. For this purpose, a ray-tracing model was implemented to generate binaural impulse responses for the binaural-model analysis. The left graph of Fig. 12 depicts the floor plan together with the trajectory of the walk-through. The encircled numbers indicate the positions of the binaural- analysis examples that as are discussed below in this section. All binaural room impulse responses for the simulations were rendered using the ray-tracing model that was discussed in Sect. 2.2. A geometrical model was defined as shown in Fig. 12b, namely, based on

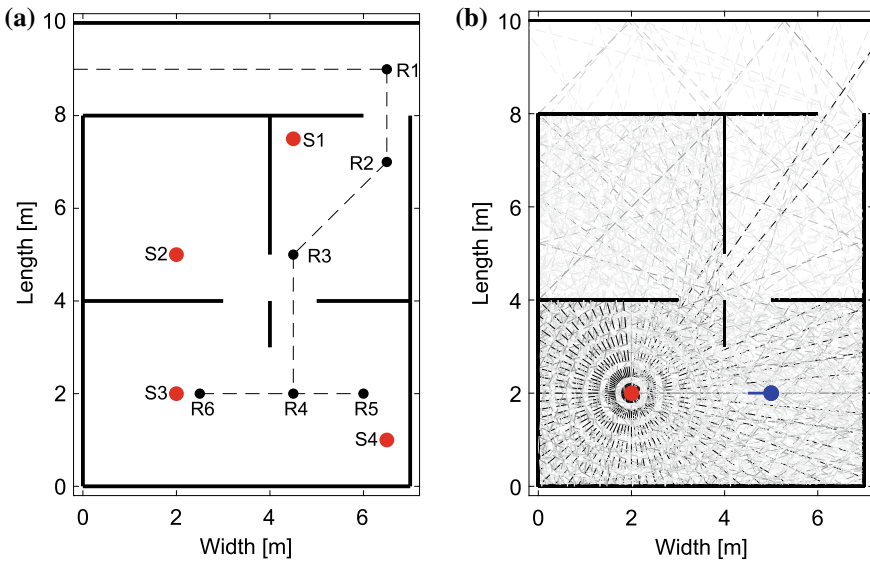


Fig. 12 **a** Diagram for the office walk-through with test positions for sources S1–S4, and receivers, R1–R6. **b** Ray-tracing simulation in a computer-generated office suite with a non-occluded sound source. The sound source is depicted as a **red dot**, the binaural receiver as a **blue dot**. The **gray level** of the rays lighten with decreasing distance and amplitude

sound-reflecting walls, a source (red dot) and a receiver (blue dot). A set of rays is sent out from the sound source at a resolution of 1 ray per 5° . Each ray is then traced, and every time a ray meets a wall it is reflected back using Snell's law, that is, considering that the outgoing angle equals the incoming angle. The ray is traced until the 20th reflection occurs unless the ray exits the geometrical model. At every reflection, the sound level is attenuated by 2 dB across frequency to simulate the acoustic absorption of the walls. The sound intensity is also attenuated over distance, based on the inverse-square law, assuming the sound source to be of omnidirectional character. The collection of rays is shown in Fig. 12b as gray lines, such that the rays become lighter in color with distance and decreasing sound pressure.

All rays are finally collected at the receiver position, assuming a spatial window of 0.6 m width. Each calculated ray is tested for whether it intersects the spatial window at the receiver position. How far each ray traveled from the source position to the receiver is then calculated for each intersecting ray. Similarly, the azimuth of the arriving ray and the order of reflection for the incoming ray is determined. Based on these data, a binaural room impulse response is calculated in which a left/right HRTF pair is inserted at the correct delay, further, the head orientation-based direction-of-arrival angle of the respective ray. Each HRTF pair is calibrated to the amplitude that the ray should have, based on the distance traveled and the number of wall reflections that it has undergone. In addition, a late-reverberation tail is generated at a constant level by assuming a statistically-evenly distributed diffuse-reverberation field, using

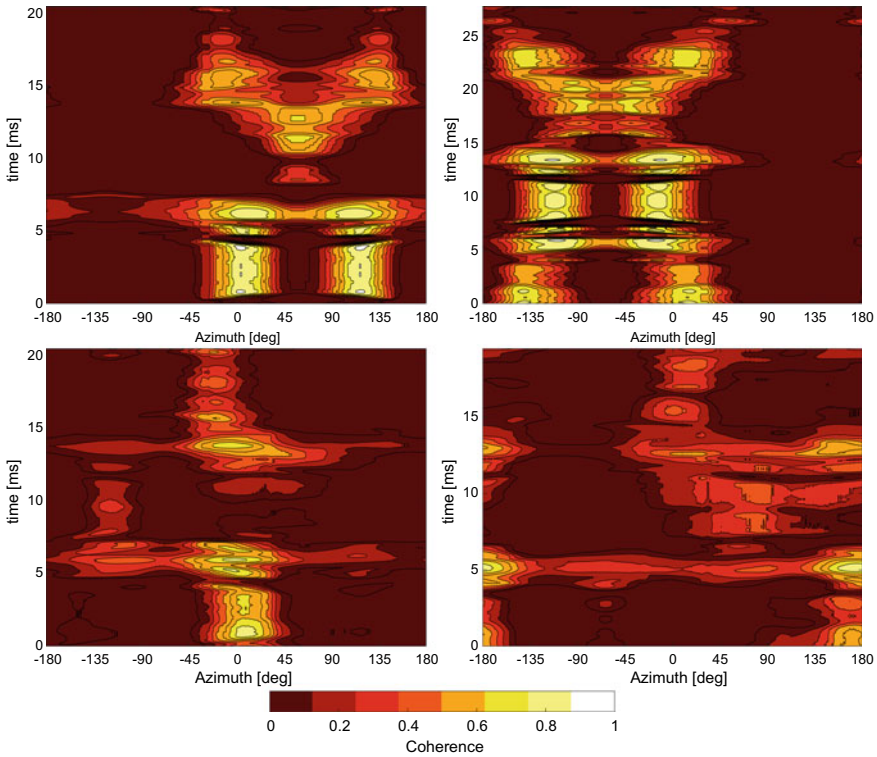


Fig. 13 Binaural-activity map results for the BICAM model analysis utilizing head movements. The **top-left graph** shows the results for Scenario 1 (Fig. 12) right, with the sound source pointing 30° left to the sound source (including a 30° head-movement compensation). The **top-right graph** shows the same condition but for the receiver pointing 30° to the right. The **bottom-left graph** shows the combined analysis for removal of front/back confusions for a receiver pointing into the direction of the sound source, 0°. The **bottom-right graph** shows the same condition as depicted in the **bottom-left graph** but for a receiver pointing away from the sound source, 180°

an exponentially decaying Gaussian noise burst adjusted to a reverberation time of 0.7 s. At the position shown in the right graph of Fig. 12, the diffuse reverberation level was about -10 dB lower than the combined level of the direct sound and the early reflections.

The results are then analyzed using the BICAM precedence-effect model (Braasch 2016) and a male-speech sample (Bang & Olufsen 1992). The BICAM algorithm was modified to transform the model’s ITD estimates into azimuths using a remapping function according to Braasch et al. (2013)—as shown in the binaural-activity map of Fig. 13 (top-left graph). The plot shows the scenario in which the virtual head of the model is turned 30° away from the sound source, based on the scenario shown in the right graph of Fig. 12. Note that the data are presented in a room-coordinate system that faces the sound source directly. As can be easily seen, each time slice shows

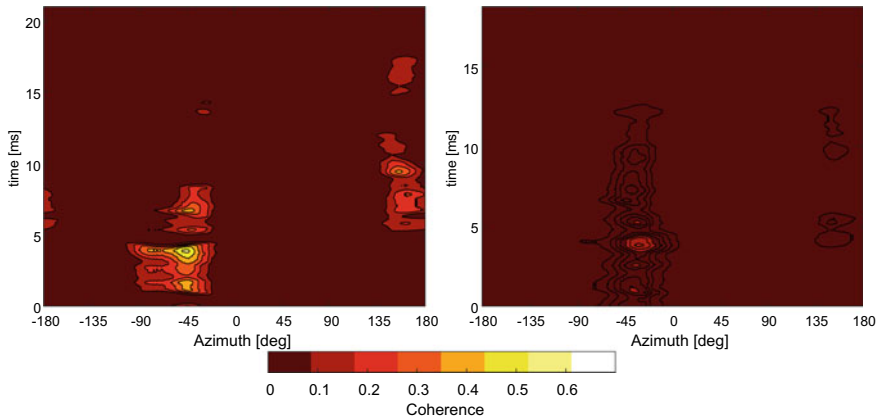


Fig. 14 Binaural-activity-map results for the BICAM model analysis utilizing head movements for an occluded direct sound source—simulating a scenario as depicted in Fig. 12a—with source position, S2, and receiver position, R6. The **left graph** shows the result for a single time interval, the **right graph** depicts the outcome for an average over 10 time intervals

two ambiguous peaks, namely, one for the front and one for the corresponding rear direction, a common problem that was discussed in detail in Braasch et al. (2013). In order to resolve the ambiguous peaks, the virtual head of the model is shifted by 60° to the opposite side—see the top-right graph of Fig. 13. This graph also displays the data in a room-related coordinate system. Now simply, the average is taken of the two binaural activity maps and the ambiguous front/back-confusion peaks average out—see the bottom-left graph in Fig. 13. To demonstrate the effectiveness of the head-movement algorithm, the same scenario was simulated again, but this time with the virtual head facing the rear at 180° with temporal head-movement shifts to 150° and 210° to resolve front/back directions. It should be noted that there are two main differences between the model presented here and the model of Braasch et al. (2013). Firstly, in the new model, the head-movement algorithm is now applied to the estimated binaural-activity map and not to the binaural signal itself. This renders to two advantages, that is, the direct-sound-source angle can be computed separately from the early reflections, which yields in a higher localization accuracy, and the algorithm can also estimate the front/back direction of the reflections. However, the new model cannot yet calculate front/back directions from a continuously turning head like it is the case for the Braasch et al. (2013) model. The reason for this is that the time-alignment method for the two autocorrelation functions currently requires a stable head orientation. Therefore, the new model calculates the front/back directions based on two distinct head positions until a better solution is found for the time alignment.

The analysis is concluded by computing a scenario in which the direct pathway between the source and the received is occluded by a wall—as shown in Fig. 12a (S2, R6). Figure 14 shows the binaural-activity maps for this case. In the left graph, the

binaural activity map was calculated from a single time interval. Here, a prominent peak is visible with a maximum correlation of 0.6 even though the direct signal was occluded. However, if the binaural-activity map is calculated as an average over 10 estimates, computed over 10 time intervals, the coherence drops to 0.2—see the right graph. The reason is that the prominent peak develops randomly at different positions for each of the ten computations. It should be noted that each segment by itself leads to a maximum coherence of one because the autocorrelation peaks always have a main peak of one. However, in the occluded case, the outcome of the analysis is heavily influenced by the diffuse-reverberant-signal component and the main peak averages out since its lateral position moves from segment to segment. In the case of Scenario 1, the binaural-activity map is stable from segment to segment and hardly influenced by the time-averaging method.

Following Wallach (1939, 1940), the head-movement model can also be used to estimate the elevation of sound sources by utilizing the fact that the ITD range is reduced with up- or downward elevation changes of the sound source from the horizontal plane—compare Pastore et al. (2020, this volume). Ideally, the ITD range is reduced monotonically with the elevation magnitude, minimizing to ITDs of zero at the -90° and 90° elevation poles.

In order to enable sound-source-elevation estimates, the head-movement model is slightly modified for processing different elevations from -70° to 80° in steps of 10° . For each elevation, a new set of frequency-dependent remapping functions is calculated according to (21). In principle, the model is an alternative implementation to an existing localization model by Parks (2014), which also draws from Wallach's ideas to estimate elevation angles. The results of the model simulation are shown in Fig. 15. Each horizontal color sequence corresponds to one elevation set as indicated on the y-axis. The sequence depicted for the elevation of 0° basically shows the same data as Fig. 11 but for different source and head-movement angles. The left graphs indicates the start position of the head-movement angle, the right graphs presents an averaged function over the head motion. The top row shows the results for a sound source at 30° azimuth and 60° elevation. At the beginning of the head-movement trajectory, the results are still ambiguous and the source could be at various elevations at 30° or 150° azimuth. After the head-movement, the model accurately locates the sound source at 30° azimuth and 60° elevation. Also in cases of a sound source located at 0° azimuth/ 0° elevation or one located at -40° azimuth/ -135° elevation, the actual sound source can be determined through head movement—see Fig. 15 center and bottom rows.

5 Conclusion

The goal of this chapter was to examine how auditory systems utilize binaural mechanisms to extract useful information from the environment to be able to read and understand a complex scene. Using idealized but complex simulated environments, it is tested how the auditory system can adapt to different scenarios. Most of the

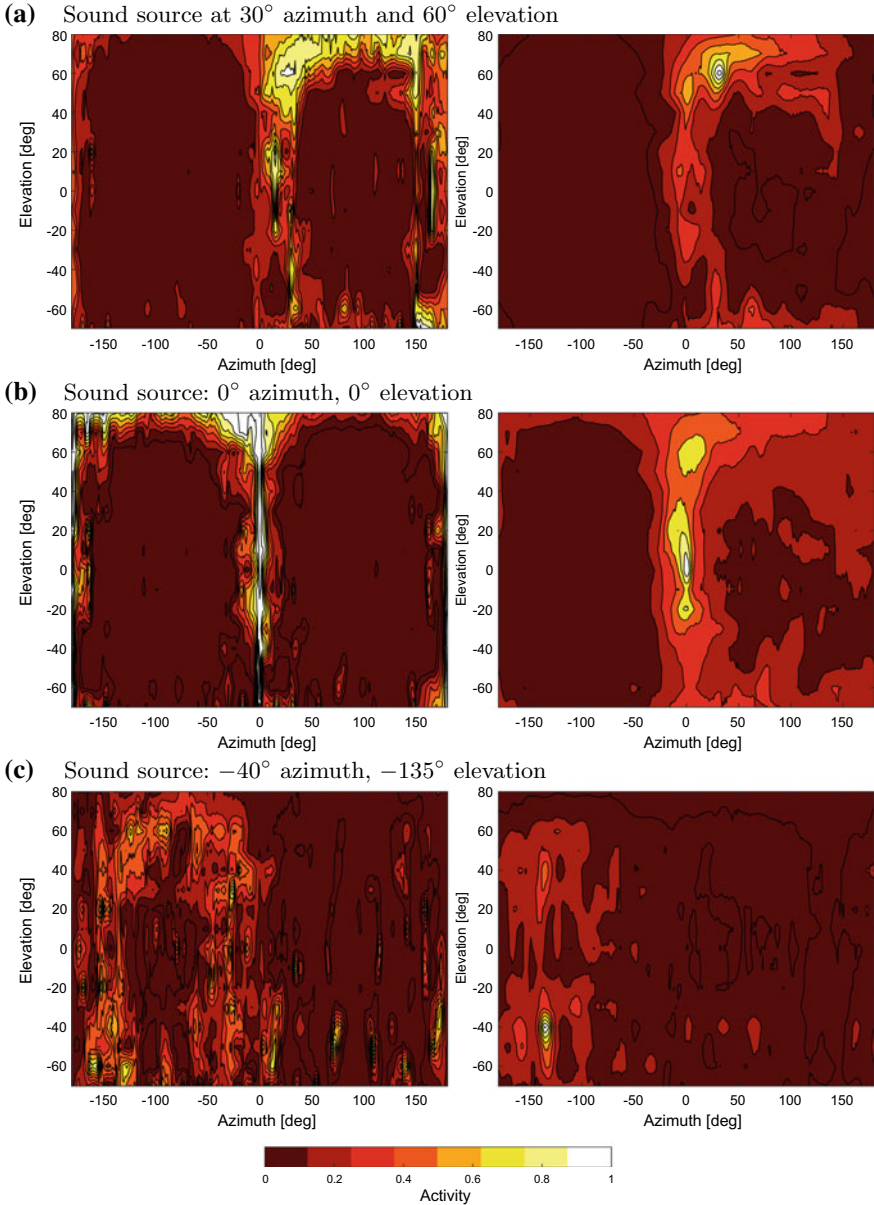


Fig. 15 Demonstration of the head-movement algorithm to estimate elevation. All **left graphs** depict the initial model performance before head movement, the **right graphs** the model performance with integrated head movement. In all cases the head rotated from 0° to 60° azimuth, maintaining the elevation at 0°. **Light areas** indicate a high likelihood of estimated source position, **dark brown areas** indicate a low probability of the source being present

auditory system's capabilities can be traced back to tasks that preceded human civilization, but a remaining mystery is how the auditory system can process specular reflections. One explanation is that our precedence mechanism is an evolutionary response to floor reflections and reflections from a cliff, where large plain surfaces can be found. Yet, it, is still amazing that this mechanism can handle sound perception in human-built rectangular rooms, which appeared very late during our evolutionary process. An alternative explanation is that the precedence effect largely falls out of the specific processing of the auditory system and is not necessarily the product of any precedence-effect specific mechanisms at all. For other tasks, the current demands are not that different from our pre-civilization experiences. Head movements, for example, can help to resolve front/back ambiguities for sound sources. A future goal is to extend the binaural analysis to actually measured environments and to support the findings with psychoacoustic experiments. The study presented here hopefully serves as an initial gateway to better understand how the binaural system reads the world under complex conditions.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant Nos. 1320059 and BCS-1539276 in alliance with the TwoEars! project (European Research Council, STREP, FP7-ICT-2013-C, No. 618075). The research was carried out within Rensselaer Polytechnic Institute's Cognitive and Immersive Systems Laboratory, a collaboration of RPI and IBM. Reiner Blumentritt provided access to the Hohle-Fels cave for the impulse-response measurement. M. Torben Pastore, William Yost, and Yi Zhou internally reviewed the manuscript. The author is appreciative for valuable comments and suggestions provided by two anonymous reviewers.

References

- Bang & Olufsen. 1992. Music for Archimedes. *CD B&O 101*.
- Baril, D. 2012. The success of homo sapiens may be due to spatial abilities. *PhysOrg*. May 09, 2012, <https://phys.org/news/2012-05-success-homo-sapiens-due-spatial.html> (last access: Dec. 15, 2019).
- Blauert, J. 1969/1970. Sound localization in the median plane. *Acta Acustica united with Acustica* 22: 205–213.
- Blauert, J. 1997. *Spatial Hearing*. Cambridge, MA: MIT Press.
- Blauert, J., and W. Cobben. 1978. Some consideration of binaural cross correlation analysis. *Acta Acustica united with Acustica* 39: 96–104.
- Blauert, J., H. Lehnert, J. Sahrhage, and H. Strauss. 2000. An interactive virtual-environment generator for psychoacoustic research. I: Architecture and implementation. *Acta Acustica United with Acustica* 86 (1): 94–102.
- Bodden, M. 1993. Modeling human sound source localization and the cocktail-party effect. *Acta acustica (Les Ullis)* 1: 43–55.
- Bonzai, M. 2018. George Massenburg builds a Blackbird room. *Digizine, Digidesign*. http://www2.digidesign.com/digizine/dz_main.cfm (last accessed: August 13, 2018).
- Braasch, J. 2002. Localization in the presence of a distracter and reverberation in the frontal horizontal plane: II. Model algorithms. *Acta Acustica United with Acustica* 88 (6): 956–969.

- Braasch, J. 2003. Localization in the presence of a distracter and reverberation in the frontal horizontal plane: III. The role of interaural level differences. *Acta Acustica United with Acustica* 89 (4): 674–692.
- Braasch, J. 2005. Modelling of binaural hearing. In *Communication Acoustics*, ed. J. Blauert, 75–108. Berlin: Springer
- Braasch, J. 2016. Binaurally integrated cross-correlation auto-correlation mechanism (BICAM). *The Journal of the Acoustical Society of America (Express Letter)* 140 (1), EL143–EL148.
- Braasch, J. 2019. *Hyper-specializing in Saxophone Using Acoustical Insight and Deep Listening Skills*. Berlin, Heidelberg: Springer.
- Braasch, J., and K. Hartung. 2002. Localization in the presence of a distracter and reverberation in the frontal horizontal plane. I. Psychoacoustical data. *Acta Acustica United with Acustica* 88 (6): 942–955.
- Braasch, J., S. Clapp, A. Parks, T. Pastore, and N. Xiang. 2013. A binaural model that analyses aural spaces and stereophonic reproduction systems by utilizing head movements. In *The Technology of Binaural Listening*, ed. J. Blauert, 201–223. Berlin, Heidelberg, New York: Springer and ASA Press.
- Breebaart, J., S. van de Par, and A. Kohlrausch. 2001. Binaural processing model based on contralateral inhibition. I. Model setup. *The Journal of the Acoustical Society of America* 110: 1074–1088.
- Bruck, D. 1999. Non-awakening in children in response to a smoke detector alarm. *Fire Safety Journal* 32 (4): 369–376.
- Busby, K.A., L. Mercier, and R. Pivik. 1994. Ontogenetic variations in auditory arousal threshold during sleep. *Psychophysiology* 31 (2): 182–188. <https://doi.org/10.1111/j.1469-8986.1994.tb01038.x>.
- Conard, N.J., and S. Wolf. 2014. *Der Hohle Fels bei Schelklingen: Ursprung für Kunst und Musik [The Hohle Fels near Schelklingen: Source of art and music]*. Schelklingen: Museumsgesellschaft Schelklingen.
- Conard, N.J., M. Malina, and S.C. Münzel. 2009. New flutes document the earliest musical tradition in Southwestern Germany. *Nature* 460 (7256): 737–740.
- Conlon, P. 2004. Reviewed work: Native America, Indian flute songs from Comanche land by Doc Tate Nevaquaya. *The World of Music* 46 (3): 208–210.
- Coolidge, F.L., and T. Wynn. 2018. *The Rise of Homo Sapiens: The Evolution of Modern Thinking*. Oxford: Oxford University Press.
- DeLoach, A.G., J.P. Carter, and J. Braasch. 2015. Tuning the cognitive environment: Sound masking with “natural” sounds in open-plan offices (A). *The Journal of the Acoustical Society of America* 137 (4): 2291.
- Deshpande, N., and J. Braasch 2017. Blind localization and segregation of two sources including a binaural head movement model. *The Journal of the Acoustical Society of America* 142 (1): EL113–EL117. <https://doi.org/10.1121/1.4986800>.
- DIN. 1968. Schallabsorptionsgrad-Tabelle (Table of absorption coefficients). In *German Standards* (Deutsches Institut für Normung (DIN)). Note: This table is no longer formally standardized.
- Ferrill, A. 2018. *The Origins of War: From the Stone Age to Alexander the Great*. Abingdon: Routledge.
- Fitch, W. 2000. The evolution of speech: A comparative review. *Trends in Cognitive Sciences* 4 (7): 258–267. [https://doi.org/10.1016/S1364-6613\(00\)01494-7](https://doi.org/10.1016/S1364-6613(00)01494-7).
- Gaudzinski-Windheuser, S., E.S. Noack, E. Pop, C. Herbst, J. Pflöging, J. Buchli, A. Jacob, F. Enzmann, L. Kindler, R. Iovita, M. Street, and W. Roebroeks. 2018. Evidence for close-range hunting by last interglacial Neanderthals. *Nature Ecology & Evolution* 1087.
- Gerkema, M.P., W.I.L. Davies, R.G. Foster, M. Menaker, and R.A. Hut. 2013. The nocturnal bottleneck and the evolution of activity patterns in mammals. *Proceedings of the Royal Society B: Biological Sciences* 280 (1765): 20130508. <https://doi.org/10.1098/rspb.2013.0508>.
- Graham, P.J. 2014. The function of perception. *Virtue Epistemology Naturalized*, 13–31. Berlin: Springer.

- Hartung, K., J. Braasch, and S.J. Sterbing. 1999. Comparison of different methods for the interpolation of head-related transfer functions. In *AES 16th International Conference on Spatial Sound Reproduction*, 319–329.
- Huang, J., N. Ohnishi, X. Guo, and N. Sugie. 1999. Echo avoidance in a computational model of the precedence effect. *Speech Communication* 27: 223–233.
- Jones, B.C., D.R. Feinberg, L.M. DeBruine, A.C. Little, and J. Vukovic. 2010. A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour* 79 (1): 57–62.
- Joris, P.X., and L.O. Trussell. 2018. The calyx of held: A hypothesis on the need for reliable timing in an intensity-difference encoder. *Neuron* 100 (3): 534–549.
- Knight, C., and J. Lewis. 2017. Wild voices: Mimicry, reversal, metaphor, and the emergence of language. *Current Anthropology* 58 (4): 435–453.
- Kuttruff, H. 2000. *Room Acoustics*. London, New York: Spon Press.
- Lehnert, H., and J. Blauert. 1992a. Aspects of auralization in binaural room simulation. In *Audio Engineering Society Convention 93*, Audio Engineering Society, Preprint No. 3390.
- Lehnert, H., and J. Blauert. 1992b. Principles of binaural room simulation. *Applied Acoustics* 36 (3–4): 259–291.
- Litovsky, R.Y., H.S. Colburn, W.A. Yost, and S.J. Guzman. 1999. The precedence effect. *The Journal of the Acoustical Society of America* 106: 1633–1654.
- Meller, H., and M. Schefzik. 2015. *Krieg: Eine archäologische Spurensuche [War: An archaeological search of traces]* (Landesamt Für Denkmalpflege und Archäologie, Landesmuseum für Vorgeschichte, Sachsen-Anhalt), Begleitband zur Sonderausstellung im Landesmuseum für Vorgeschichte, Halle an der Saale, Germany, 6 Nov 2015–22 May 2016.
- Meyer, J. 1978. Raumakustik und Orchesterklang in den Konzertsälen Joseph Haydns [room acoustics and orchestra sound in Joseph Haydn's concert halls]. *Acta Acustica united with Acustica* 41 (3): 145–162.
- Mi, J., M. Groll, and H.S. Colburn. 2017. Comparison of a target-equalization-cancellation approach and a localization approach to source separation. *The Journal of the Acoustical Society of America* 142 (5): 2933–2941. <https://doi.org/10.1121/1.5009763>.
- Parks, A.J. 2014. Listener orientation and spatial judgments of elevated auditory percepts. Ph.D. thesis, Rensselaer Polytechnic Institute.
- Pastore, M., Y. Zhou, and W.A. Yost. 2020. Cross-modal and cognitive processes in sound localization. In *The Technology of Binaural Understanding*, eds. Blauert, J. and J. Braasch, 315–350. Cham, Switzerland: Springer and ASA Press.
- Patterson, R.D., M.H. Allerhand, and C. Giguère. 1995. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America* 98 (4): 1890–1894.
- Peck, J.E. 1994. Development of hearing. Part I: Phylogeny. *Journal American Academy of Audiology* 5: 291.
- Peck, J.E. 1995. Development of hearing. Part III: Postnatal development. *Journal American Academy of Audiology* 6: 113.
- Reethof, G., O. McDaniel, and G. Heisler. 1977. Sound absorption characteristics of tree bark and forest floor. In *Proceedings of the Conference on Metropolitan Physical Environment*, eds. L.P. Heisler, M. Gordon and L.P. Herrington, 206–217. Gen. Tech. Rep. NE-25. Upper Darby, PA: U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station.
- Reznikoff, I. 2004/2005. On primitive elements of musical meaning. *The Journal of Music and Meaning* 3. www.musicandmeaning.net/issues/showArticle.php?artID=3.2 (last accessed: July 2017).
- Roebroeks, W., N.J. Conard, T. Van Kolfschoten, R. Dennell, R.C. Dunnell, C. Gamble, P. Graves, K. Jacobs, M. Otte, D. Roe, et al. 1992. Dense forests, cold steppes, and the palaeolithic settlement of Northern Europe - [and comments and replies]. *Current Anthropology* 33 (5): 551–586.
- Roman, N., D. Wang, and G.J. Brown. 2003. Speech segregation based on sound localization. *The Journal of the Acoustical Society of America* 114 (4): 2236–2252.

- Roman, N., S. Srinivasan, and D. Wang. 2006. Binaural segregation in multisource reverberant environments. *The Journal of the Acoustical Society of America* 120 (6): 4040–4051.
- Rönne, H. 1915. Zur Theorie und Technik der Bjerrumschen Gesichtsfelduntersuchung [On the theory and technique of Bjerrum's visual field examination]. *Archiv für Augenheilkunde* 78: 284–301.
- Sakai, H., S.-I. Sato, and Y. Ando. 1998. Orthogonal acoustical factors of sound fields in a forest compared with those in a concert hall. *The Journal of the Acoustical Society of America* 104 (3): 1491–1497.
- Sala, N., J.L. Arsuaga, A. Pantoja-Pérez, A. Pablos, I. Martínez, R.M. Quam, A. Gómez-Olivencia, J.M.B. de Castro, and E. Carbonell. 2015. Lethal interpersonal violence in the middle pleistocene. *PLoS one* 10 (5): e0126589.
- Sayers, B.M., and E.C. Cherry. 1957. Mechanism of binaural fusion in the hearing of speech. *The Journal of the Acoustical Society of America* 29: 973–987.
- Stern, R.M., and H.S. Colburn. 1978. Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position. *The Journal of the Acoustical Society of America* 64: 127–140.
- Teret, E., M.T. Pastore, and J. Braasch. 2017. The influence of signal type on perceived reverberance. *The Journal of the Acoustical Society of America* 141 (3): 1675–1682.
- Villa, P., and S. Soriano. 2010. Hunting weapons of Neanderthals and early modern humans in South Africa: Similarities and differences. *Journal of Anthropological Research* 66 (1): 5–38.
- Volf, M. 2012. Effects of soundscapes on attentional capacity. Master's thesis, Rensselaer Polytechnic Institute, Troy, NY.
- Vorländer, M. 1989. Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm. *The Journal of the Acoustical Society of America* 86 (1): 172–178.
- Wallach, H. 1939. On sound localization. *The Journal of the Acoustical Society of America* 10 (4): 270–274.
- Wallach, H. 1940. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology* 27 (4): 339–368.
- Wightman, F.L., and D.J. Kistler. 1992. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America* 91: 1648–1661.
- Wynn, T., and F.L. Coolidge. 2004. The expert Neandertal mind. *Journal of human evolution* 46 (4): 467–487.
- Wynn, T., and F.L. Coolidge. 2008. A stoneage meeting of minds: Neandertals became extinct while Homo Sapiens prospered. A marked contrast in mental capacities may account for these different fates. *American Scientist* 96 (1): 44–51.
- Zakarauskas, P., and M.S. Cynader. 1993. A computational theory of spectral cue localization. *The Journal of the Acoustical Society of America* 94 (3): 1323–1331.

Processing Cross-Modal Inference

Psychophysical Models of Sound Localisation with Audiovisual Interactions



Catarina Mendonça

Abstract Visual signals can have an important impact on the perceived location of sound sources. Neurological mechanisms enable interactions between seeing and hearing to form a sense of space. The effect of vision on auditory localisation percepts is of fundamental importance. A sound source is either perceived at the location of the visual source or it is perceptually shifted toward its direction. This bias is one form of visual capture. The extent of the interactions depends on time and space constraints beyond which visual and auditory cues do not necessarily interact. These constraints and interactions vary for the localisation of sources along the horizontal and vertical planes, as well as with distance. While the traditional models of audiovisual interaction in space perception assume sensory integration, recent models allow for sensory cues to either interact or not. Models of visual dominance, modality appropriateness, and maximum likelihood estimation predict one combined percept. The newer models of causal inference allow for varied perceptual outcomes depending on the relationship of the different sensory cues. Finally, visual spatial cues can induce changes to how sounds are localised after the audiovisual experience. This notorious effect, known as the *ventriloquism aftereffect*, is possibly the main mechanism of auditory space learning and calibration. The *ventriloquism aftereffect* has been described with a causal inference model and with an inverse model. The current chapter discusses all of the above concepts, establishing a connection between psychophysical data and available models.

1 Introduction

The understanding of auditory space perception mechanisms is incomplete without considering audiovisual interactions. Very rarely, in daily life, does one localise sounds without any source of visual information available. Vision is known to affect

C. Mendonça (✉)

Center for Psychology, Faculty of Psychology and Education Sciences, University of Porto,
R. Alfredo Allen, 4200-135 Porto, Portugal
e-mail: ana.cm.hipakka@uac.pt

© Springer Nature Switzerland AG 2020

J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_11

289

largely where in space sound is perceived. When conditions are met, vision can be so powerful that sounds are perceived to be sourced in the same position as the visual stimulus source. Throughout the years, the understanding of the psychophysical interactions between sound and light stimuli developed, and so did the proposed mechanisms and models. Most of the original models were not expressed mathematically, probably due to their simplicity. The newer models are formulated mathematically, and their complexity and comprehensiveness have expanded considerably. Understanding and comparing these models is helpful to comprehend perceptual mechanisms associated with the localisation of multisensory events.

Here, the evolution of these models throughout time is described, connecting them to known psychophysical effects. First, a summary of psychophysical findings is presented that describe sound localisation in the presence of audiovisual stimulation. This section analyses under which conditions light and sound stimuli interact in the formation of auditory space. Neural substrates and localisation in azimuth, elevation, and distance are briefly described. The changes in localisation of sound sources after exposure to audiovisual events are also summarised.

The remainder of this chapter highlights psychophysical localisation models that include multisensory interactions. The chapter starts with models that assume cue integration and later present models that do not assume such integration. Here, *assuming integration* refers to whether one or two percepts are formed. When there is a multisensory event where sound and light sources do not match in space, the resulting percept can be of only one unified event, in which the source positions may be wrongly localised in order to match each other; or there can be two independent percepts, one for each sensory modality. As will be seen below, the original theories of multisensory interaction in spatial perception tended to assume that there would be integration. More recent models account for the fact that stimuli might or not be integrated.

The final part of this chapter presents two attempts to model sound localisation after audiovisual experience. Audiovisual experience is known to affect the subsequent ability to localise sounds, and it is a prominent source of auditory-space learning and calibration. Modelling these sensory learning processes remains an open challenge.

All models presented in this chapter are psychophysically motivated. This means that they establish a mathematical relationship between external physical and internal psychological quantities. In all cases, these models establish relationships between the location of the external sources and their perceived positions.

The models described in this chapter can be applied directly to multimedia technology and sensor fusion. These models have been applied to automatically identifying sources from combined video and audio data. When combining sensor data, quite often information from different sensors will be incongruent. The task of identifying the degree of compatibility, which signals to rely on most, and how to combine the signals is not unlike the perceptual tasks humans face daily. As will be shown, human perception is hugely optimised. Sensory processing seems to deal optimally

with signal noise. Therefore, understanding the mechanisms of signal integration in humans might provide a good starting point for the design of technology that merges visual and audio signals.

2 Summary of Findings on Sound Localisation from Audiovisual Stimulation

2.1 Neural Substracts

Several brain regions are known to be involved in the processing of multisensory interactions. Among them, the Superior Colliculus (SC) in the midbrain seems to have particular relevance. Signals from visual, auditory and somatosensory areas converge in the SC (Meredith and Stein 1986). Along the auditory pathway, signals reach the inferior colliculus before they reach the auditory cortex. Further, there are bi-directional projections between the auditory cortex and inferior colliculus. The relationship between the inferior and superior colliculi may drive processes of spatial attention and spatial learning. It has been argued that the SC contains an amodal representation of space, which might be responsible for most multisensory interactions in source localisation (Hartline et al. 1995; Wallace et al. 1996). The cells in the mammalian SC react according to rules that resemble common rules of multisensory interaction. Given two stimuli from different sensory modalities, these cells respond stronger when both stimuli are weaker than when one stimulus is substantially stronger than the other. This effect is known as inverse effectiveness. The cells in the SC also respond more strongly when stimuli from both sensory modalities are co-localised in space, and when they occur synchronously. Chapter 6 discusses in greater detail the biological aspects of perceptual auditory space formation.

2.2 Localisation in Azimuth

Throughout the years, research has revealed rules and constraints to describe the interactions between light and sound processing in the perception of space. In the presence of visual stimulation, the task of perceptually localising sounds in azimuth is affected. If the audio-visual stimulation is matching, there are less errors and greater precision in sound localisation. If audio-visual sources are not matching, there tend to be errors in sound source localisation in order to match the visual stimulation. The three main constraints to the degree of audio-visual match are congruence, temporal matching, and spatial matching. These constraints have mostly been studied for azimuthal sound-source estimation tasks.

The *constraint of congruence* describes how the visual and the auditory signal relate to each other semantically or in terms of context. This factor is critical for multisensory integration, and some degree of congruence is required for interactions to be observed (Laurienti et al. 2004). For instance, the sound of football playing presented synchronously with images of a falling tree might evoke weaker audiovisual interactions than when it is presented with congruent images of a football player.

The constraint of temporal matching relates to the timing of the audiovisual events. When visual and auditory events occur synchronously or within a given time window, they may interact in the formation of percepts. The visual stimulus can alter the perceived sound source location if the delay between both stimuli is less than 100 ms, and the sound stimulus arrives after the visual stimulus. The temporal window within which there are multisensory interaction effects is also known as the “integration window”. Note that this window of integration is somehow related to the window of perceived simultaneity between visual and auditory signals. However, the window of integration is narrower than the window of perceived simultaneity (Slutsky and Recanzone 2001). This indicates that stimuli must be perceived as unambiguously synchronised if they are to interact in the formation of the space percept. It must be noted that the window of perceived audiovisual simultaneity varies across task, stimuli, and even number of present spatial cues (Van Eijk et al. 2008; Silva et al. 2013).

The constraint of spatial matching relates to how close in space audiovisual events occur. It has been suggested that a maximum of 15° of separation was needed for sounds to be perceived in the same location as the image (Slutsky and Recanzone 2001). However, interactions between the signals can occur even if they are not perceived as co-localised. There are biases in the localisation of sound sources with spatial discrepancies between light and sound of up to 25° (Bertelson and Radeau 1981; Wozny and Shams 2011). This bias is observed in the form of a displacement of the perceived sound source location, in magnitudes ranging between 4° and 8.2° for source separations from 7° to 25°, respectively. The further away the stimuli are from each other, the larger the average displacement of perceived sound location. Mendonça et al. (2015) observed that the visual stimuli were usually perceived accurately in space when visual and sound sources were separated by 12°. The sound source moved perceptually toward the visual source, although they were not co-localised—see Fig. 1. As will be discussed further below, it is still uncertain in this case if the subjects perceive the sound as displaced, or if they perceive it as alternating between stemming from the sound source and from the visual source.

The interactions between light and sound in the perceived azimuthal source location can be summarised as follows. The visual stimulus can attract the auditory percept, but it is debatable whether there is an influence of the auditory input on the visual percept. These interactions are better described, tested, and understood when applying psychophysical models, as described in Sect. 3.

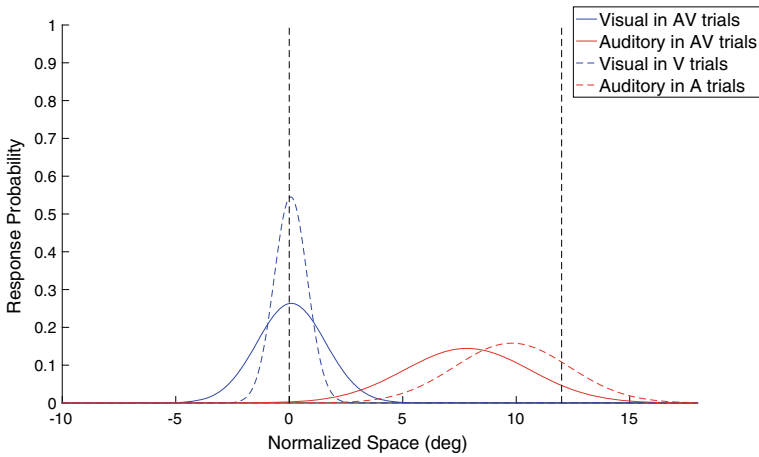


Fig. 1 Data from a localisation experiment where visual and audio signals were presented 12° apart in discrepant trials (Mendonça et al. 2015). The visual stimulus was localised very accurately in unimodal visual trials (Visual in V trials) and in the audiovisual discrepant trials (Visual in AV trials), although with lower precision. The auditory stimulus was localised poorly in both the unimodal (Auditory in A trials) and audiovisual trials (Auditory in AV trials). There was a noticeable displacement of the auditory percept in the audiovisual trials that generalised to the unimodal trials

2.3 Localisation in Elevation

To date, there have been few contributions to the understanding of how visual and audio signals interact in the perception of vertical auditory space. It cannot be assumed that the multisensory effects will be similar to those of horizontal sound localisation, as described in Sect. 2.2, since the auditory system uses different cues to localise in elevation. There is also lower precision for vertical sound-source localisation, compared to horizontal localisation, and there are constraints to audiovisual interactions, such as minimum stimulus duration and field of view.

Werner et al. (2013) created sounds changing in elevation by using individual head-related-transfer-function measurements. They used these sounds, reproduced through headphones, to test the influence of light on the perceived location of sound sources. It was found that audiovisual sources must be positioned within 7° to 10° of each other in elevation so to be perceived as co-located. Furthermore, when displaced, the visual stimulus is able to shift the perceived location of the sound source. However, the observed shifts are only of a maximum magnitude of 3.2° to 3.6° with audiovisual vertical displacements of up to 30°. This effect is close to that observed in azimuth, or smaller. This finding seems to contradict the assumptions of the models discussed in Sect. 3, where greater uncertainty in sound localisation leads to greater influence of the visual cue, and therefore to a larger displacement of the spatial percept. Unfortunately, no model of audiovisual source localisation has ever been

tested for the localisation of sources varying in elevation. Therefore, more studies are needed to assert whether the models of audiovisual localisation apply to localisation in the vertical plane.

2.4 Localisation in Distance

Mendonça et al. (2016) have recently studied how light and sound signals interact in distance perception of external sources. They found that sound and light stimuli are perceived to be co-localised if they are within 1 m of each other, but no further. There is a trend to be more tolerant to stimuli mismatch with increasing distance. Stimuli that are several meters away from the observer can be more than 3 m apart from each other and still be perceived as co-localised—see Fig. 2.

The distance localisation of visual sources is more accurate than that of sound sources. When presented together with varying degrees of spatial mismatch, the visual percepts are not influenced by the presence of sound. Sound distance estimates are affected by the presence of visual sources, but the magnitude of this effect is small and requires both sources to be close in space—see Fig. 8. The spatial window within which the visual cue may affect the auditory percept is presented in Fig. 3. The interaction windows are larger than the co-localisation windows seen in Fig. 2, but they too seem to grow with stimuli distance. Interestingly, distance perception

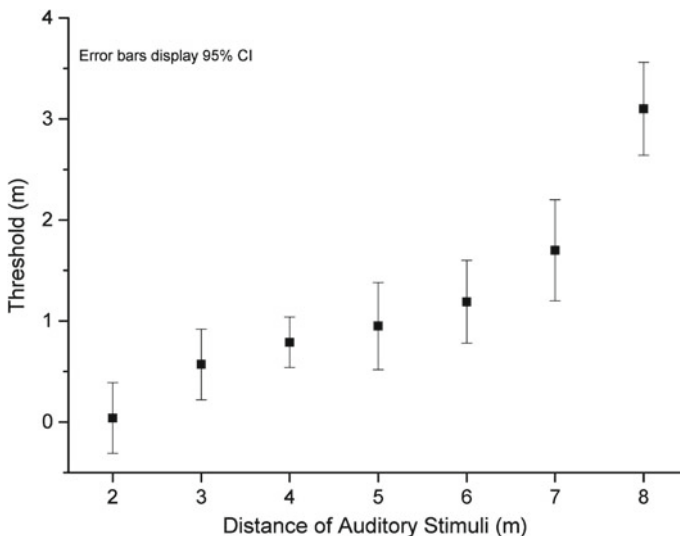


Fig. 2 The threshold of how large the distance between light and sound sources can be, in order to be perceived as co-located. The threshold denotes the point where stimuli are co-localised in 50% of the trials. The thresholds and respective 95% CI bars were determined through a bootstrapping technique

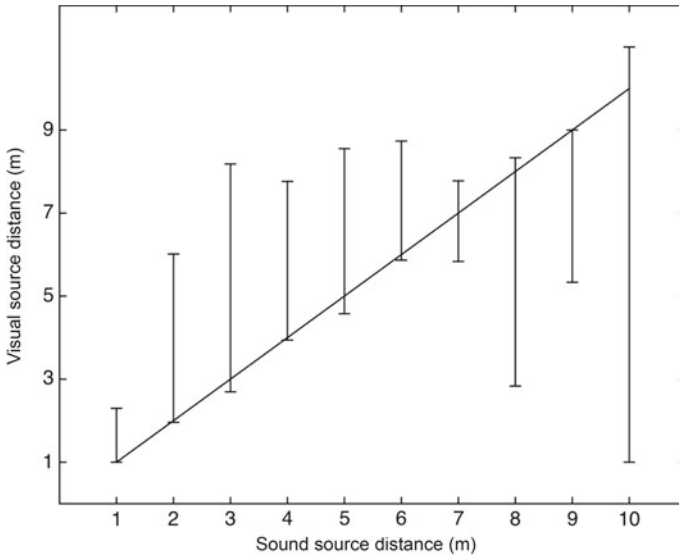


Fig. 3 Windows of interaction in the perception of auditory distance. Within this window, the visual stimulus has a weight of 50% or more on the formation of the auditory-distance estimate. Sensory weights are determined according to Eq. 21

seems to be completely unaffected by the concurrent visual cue when both stimuli are far apart. This is not the case when judging the azimuth for multimodal stimuli (e.g., Wozny and Shams 2011; Mendonça et al. 2015).

2.5 Sound Localisation After Audiovisual Stimulation

It has been established that visual cues can affect the perceived sound source position. This effect can lead to long-lasting changes. When consistently exposed to audiovisual stimuli which are mismatched in space, a recalibration mechanism can be observed whereby sound localisation is permanently shifted. This effect is thought to be one of the main mechanisms through which sound localisation is learned (King 2009). Since sound localisation cues change throughout life, due to changes in pinna shape, changes in head shape through growth, and changes in mechanical and in sensorineural sound signal processing, adaptations must take place to ensure that the ability to localise sound sources is preserved. Because of these changes, there are permanent shifts in auditory source localisation in the direction of the bias induced by the visual cue. The first study to identify this spatial aftereffect was conducted by Held (1955). The author used pseudophones. These devices function like a hearing-aid and operate by shifting the perceived sound consistently horizontally. After wearing the pseudophones in their daily lives, listeners experienced lasting changes in the

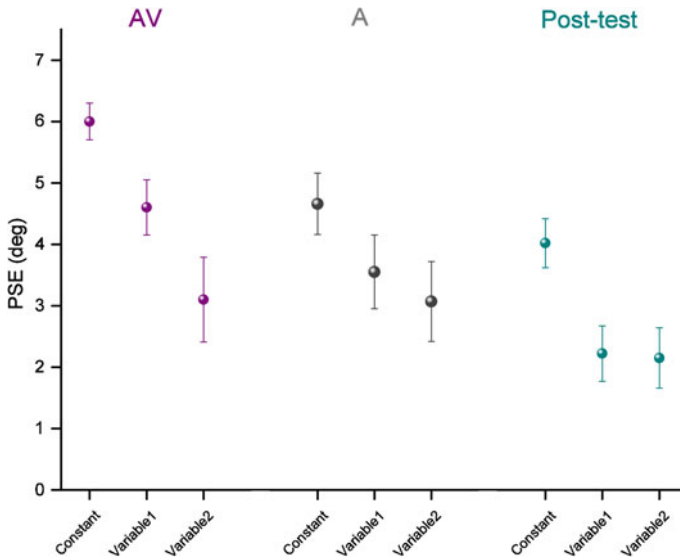


Fig. 4 Degree shift at the point of subjective equality (PSE) between presented and perceived sound-source locations. These shifts were observed during audiovisual trials (AV), auditory-only trials immediately after the audiovisual trials (A), and in auditory trials in a subsequent test session (Post-test). In the condition “Constant”, audiovisual stimuli occurred in random positions, but visual events were always 6° away from the sound event. In the condition “Variable 1”, stimuli were randomly displaced according to a Gaussian distribution with $\mu=6^\circ$. The condition “Variable 2” had a similar distribution, with a mean of 6° , but it was skewed such that the peak of the function (the mode) had a different value from the mean

perceived sound source positions. Other initial experimental studies on the ventriloquism aftereffect exposed subjects to audiovisual events that were always shifted by a few degrees for a period of 20–30 min. In after-test sessions, it was observed that subjects had a shift of a few degrees for the perceived sound location, even in absence of visual stimulation (Radeau and Bertelson 1974; Recanzone 1998).

The consistency of the stimulation is crucial in this effect. A fixed discrepancy between light and sound source positioning will lead to more pronounced subsequent shifts in sound localisation, compared to presenting subjects with audiovisual pairs of variable discrepancy levels (Mendonça et al. 2014). Figure 4 shows the effect of exposure to audiovisual stimulation with either fixed or variable discrepancy. It can be observed that the impact of the visual stimulus during the audiovisual trials is larger when the mismatch is constant, compared to the other conditions. In the post-test, the magnitude of the aftereffect is twice as pronounced in the case of consistent stimuli.

Recent research has shown that prolonged exposure is not required to obtain a shift in auditory space. It was found that after a single, brief presentation of a spatially discrepant audiovisual event, there was a measurable displacement of the perceived sound source location in the subsequent trial (Wozny and Shams 2011; Mendonça

et al. 2015). In the study by Wozny and Shams (2011) it was found that localisation of the auditory image was affected by discrepant audiovisual events that happened in the previous trial and up to three trials back. Mendonça et al. (2015) tested different models of sequential effects and found that all recent audiovisual experience as a whole has an influence on the auditory localisation estimate, but the last trial has a far higher impact than any of the preceding audiovisual events. This model is further described in Sect. 4.

3 Modelling Sound Localisation During Audiovisual Stimulation

The models described in this section are not specific to the localisation of sound in azimuth, elevation, or distance. All models have been applied to sound localisation, but they could also be applied to other multisensory processes.

3.1 Models That Assume Cue Integration

Visual Capture, Dominance and the Ventriloquism Effect

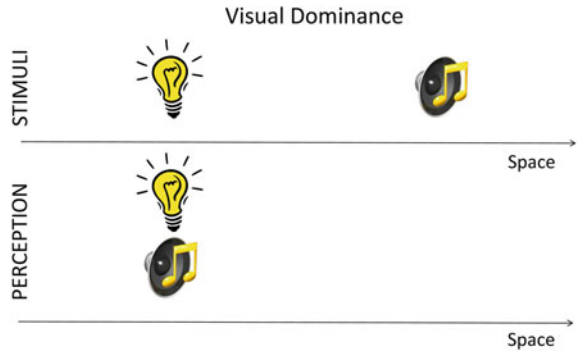
The concept of visual capture was first proposed by Rock and Victor (1964) in a study about the visual capture of touch in the perceived size of objects. This concept, also called visual dominance (Posner et al. 1976), was later found to describe well the localisation of sound from audiovisual stimulation. When both visual and auditory events are presented simultaneously and in proximity, the visual stimulus determines where the auditory stimulus is perceived (Bertelson and Radeau 1981; Choe et al. 1975; Bertelson 1999). This perceptual effect is a form of ventriloquism effect. When a ventriloquist actor speaks while moving their puppet's mouth, there is an impression that the sound is originating from the puppet itself, rather than from the actor. This represents well what is assumed to happen in the case of visual dominance for the perceived sound source location. The mechanism of visual dominance is depicted in Fig. 5.

For decades, this mechanism was assumed to describe well where sound sources are perceived in the presence of visual stimuli (e.g., Vroomen et al. 2001). Sensory dominance can be expressed conditionally, where the sensory estimate of the sound location, \hat{s}_A , given the visual source location, x_V , and the auditory source location, x_A , is the same as the sensory estimate of the visual location, \hat{s}_V :

$$\hat{s}_{A|x_V, x_A} = \hat{s}_{V|x_V, x_A}. \quad (1)$$

In this theory, the estimate of the visual source location is unaffected by the auditory modality,

Fig. 5 Schematic of the expected multisensory interactions in the perceived location of visual and auditory sources for the visual dominance model



$$\hat{s}_{V|x_V, x_A} = \hat{s}_{V|x_V}. \tag{2}$$

In most statistical models of audiovisual perception, the internal sensation, s_i , is a noisy sensory representation caused by a given stimulus, x_i , which can be approximated by the observed sensory estimates, \hat{s}_i . Here, the index i stands for each individual sensory modality. These estimates can be approximated following the Bayesian rule,

$$p(\hat{s}_i|x_i) = \frac{p(x_i|\hat{s}_i) p(\hat{s}_i)}{p(x_i)}. \tag{3}$$

The sensory estimates $\hat{s}_i|x_i$ are the posterior likelihood, which is approximated by the stimulus location, $p(x_i|\hat{s}_i)$, the prior internal perceptual bias, $p(\hat{s}_i)$, and the prior external likelihood of that stimulus being in that location, $p(x_i)$. Most modellers assume that the external world is unbiased, and therefore $p(x_i) = 1$. However, the observer can use his or her knowledge about the world to form the percept, and, therefore, sometimes $p(\hat{s}_i)$ assumes different values. In the remainder of this document, $p(\hat{s}_i)$ is referred to as p_i to conform with standard language in the field. In the case of visual dominance, the perceived location of the auditory source can be approximated as:

$$p(\hat{s}_A|x_A, x_V) = p(\hat{s}_V|x_V) = p(x_V|\hat{s}_V) p(V). \tag{4}$$

Even though the concepts of visual dominance and ventriloquist effect remain very popular in current literature, these concepts are outdated. The above model will successfully predict human sound localisation in a large number of contexts. However, it is considered that other, more comprehensive models, might better grasp the interactions in audiovisual localisation.

Modality Appropriateness and Modality Precision

The concepts of *modality appropriateness* and *modality precision* were first described many decades ago (Welch and Warren 1980). These concepts vary in the proposed mechanisms, but have the same practical implications. Modality appropriateness

assumes that some sensory modalities are more suitable than others for the perception of a given attribute, due to their different information-processing characteristics. It was, therefore, hypothesised that, when having data from both senses available, the brain would choose to use only the data from the most reliable sense. In the presence of visual-spatial information, auditory spatial information would always be neglected, because of the relatively weaker spatial resolution of the auditory system. In modality precision, it is assumed that the data from the most reliable source of sensory information will always be used. This is very similar to modality appropriateness, but it does not assume that one sensory modality will always be best, instead putting the focus on the quality of the sensory signal. Expressing this mathematically, one can state that the auditory localisation estimates are the same as the visual estimates, and what determines which sensory cue dominates is its reliability, which can be defined as lower variance of the sensory information σ_i^2 ,

$$p(\hat{s}_{A|x_V, x_A}) = p(\hat{s}_{V|x_V, x_A}) = \begin{cases} p(x_A|\hat{s}_A) p(A) & \text{if } \sigma_V^2 > \sigma_A^2 \\ p(x_V|\hat{s}_V) p(V) & \text{if } \sigma_V^2 < \sigma_A^2 \end{cases}. \quad (5)$$

This model advances from the previous one by allowing the auditory stimulus to dominate in the final estimate under sensory conditions where the visual information is less precise or has a poor resolution. The model has been compared to other recent models in estimating distance perception, and it was found to perform poorly for the prediction of auditory distance, but worked well to predict perceptual visual distance (Mendonça et al. 2016).

Maximum Likelihood Estimation

The Maximum Likelihood Estimation (MLE) model was developed by Yuille and Bulthoff (1996), extended by Landy et al. (1995), and used in modelling of multi-sensory interactions for the first time by Ernst and Banks (2002). For this model, the sensory estimates, \hat{s}_i , and their variances, σ_i^2 , are assumed to be normally distributed and independent. It is further assumed that the Bayesian prior is uniform ($p(i) = 1$). According to the maximum likelihood approach, the sensory estimate can be given by a linear weighted sum of the individual unimodal sensory estimates,

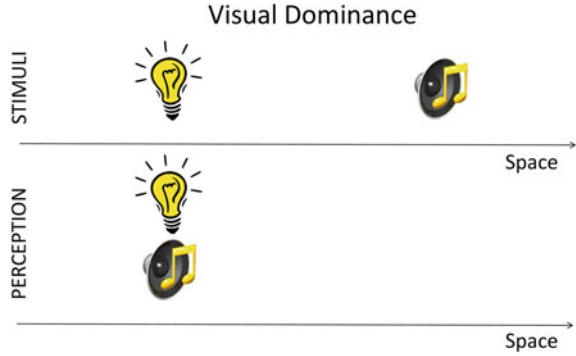
$$\hat{s} = \sum_i w_i \hat{s}_i. \quad (6)$$

In the case of audiovisual localisation this becomes

$$\hat{s}_{AV} = \hat{s}_A w_A + \hat{s}_V w_V, \quad (7)$$

where w_i are the weights of the individual sensory percepts. For the optimal estimation, the weights correspond to the inverse of the estimate's variance

Fig. 6 Schematic of the expected multisensory interactions in the perceived location of visual and auditory sources in the MLE model



$$w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2}. \tag{8}$$

Therefore, in the case of audiovisual localisation, the predicted combined variance of the final estimate is

$$\sigma_{AV}^2 = \frac{\sigma_A^2 \sigma_V^2}{\sigma_A^2 + \sigma_v^2}. \tag{9}$$

This approach is compatible with previous findings supporting the ventriloquism effect. This model suggests that visual cues tend to dominate sound source localisation because they carry a higher sensory weight. The localisation of visual events is indeed more precise and typically carries lower variances than the localisation of auditory events. This imbalance of sensory cue reliability can lead to results that resemble the visual dominance effect. Note that this model consistently merges audiovisual cues. It always produces a united percept, even if visual and auditory sources are separated in space. Figure 6 shows the perceptual mechanisms predicted by the MLE model. Spatially separated auditory and visual stimuli give rise to a fused percept. The percept is more defined in space than for any of the original signals, because the variance of the percept is optimised through the combination of information.

This model has been shown to accurately describe human audiovisual perception, in particular audiovisual localisation (Alais and Burr 2004; Binda et al. 2007). The model is still considered to be accurate, however, multisensory integration only occurs when certain conditions are met. Audiovisual interactions predicted by the MLE break down when stimuli from the different sensory modalities are incongruent or not co-localised (e.g., Mendonça et al. 2011). Therefore, there seems to be a window of cue compatibility within which multisensory integration is observed. However, the MLE model is no longer applicable beyond this window.

3.2 Models That Do Not Assume Cue Integration

No Interaction

Sometimes, it can be found that there is no perceived spatial interaction between visual and auditory cues, but this has never been proposed as a complete model on its own. For instance, when the cues are unrelated or too far apart, it can be observed that auditory localisation is unaffected by the presence of visual stimuli. In the *no-interaction* model, the perceived sound source position is solely a function of auditory cues and independent from visual cues. Therefore, the localisation of a sound source can be approximated as

$$p(\hat{s}_A|x_A, x_V) = p(x_A|\hat{s}_A) p(A), \quad (10)$$

like in (3).

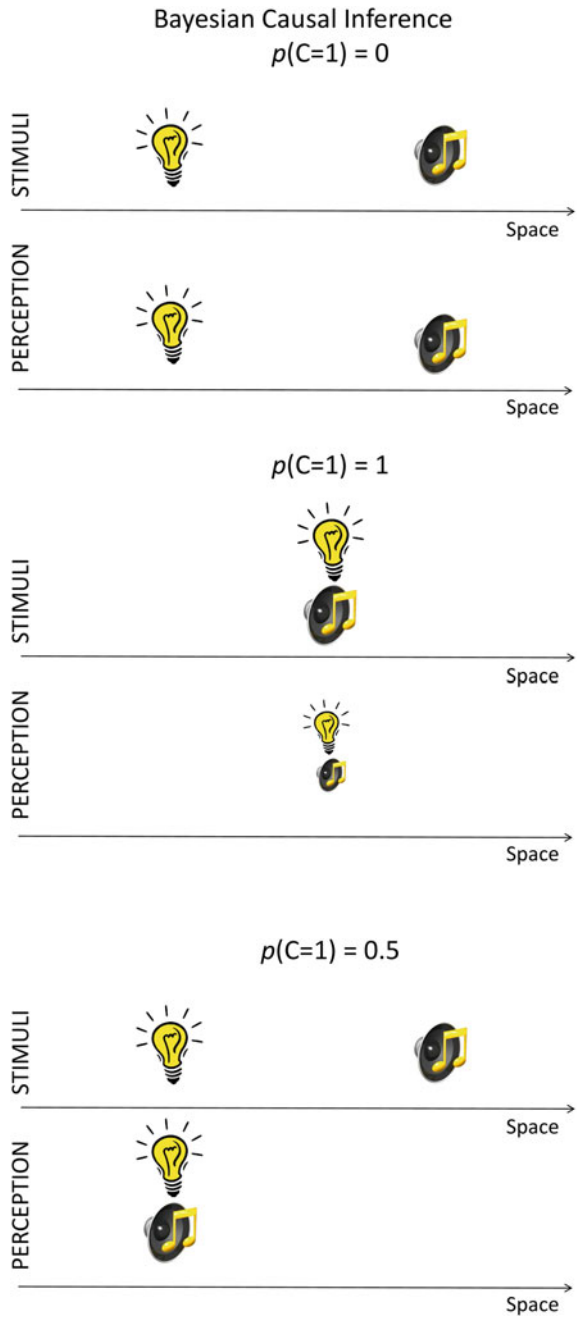
Causal Inference with a Generative Model

As a response to the limitation of the MLE model, which fails to provide an accurate prediction of the localisation of sound when visual and auditory cues are too incongruent, a new kind of model emerged. This type of model predicts optimal integration only in some instances. Among others, two approaches were proposed in multisensory research to deal with the problem of different sensory cue interaction rules according to different degrees of multisensory cue matching. In this context, Roach et al. (2006) worked on audiovisual rate perception, while Bresciani et al. (2006) focused on audiotactile integration. These models used a Gaussian ridge to quantify the binding between the different sensory modalities. The first model of this kind to predict audiovisual localisation specifically was the Bayesian causal inference model (Körding et al. 2007; Sato et al. 2007). This model specifies that there is a sequence of steps to multisensory cue processing. Unimodal stimuli are processed separately at an early stage. Then, sensory estimates are plotted against each other and the degree of compatibility is computed. From that, the brain estimates the likelihood that both estimates were caused by the same event. Figure 7 illustrates the predicted interactions between visual and auditory cues is space perception according to this model.

As shown in Fig. 7, top graph, the model predicts that when cues are too far apart they do not interact and, therefore, the percepts are unbiased by the concurrent sensory stimulation. When stimuli are close enough in space, they are integrated, and the percept becomes optimal, as predicted by the MLE model—see Fig. 7, center graph. In intermediate situations there are variable results, where stimuli can be perceived closer to each other, but are not necessarily co-localised—see Fig. 7, bottom graph. The accuracy of these estimates may change. The causal inference model follows a series of three steps. Firstly, causality is established through the Bayesian equation

$$p(C|x_A, x_V) = \frac{p(x_A, x_V|C) p(C)}{p(x_A, x_V)}. \quad (11)$$

Fig. 7 Schematic of the expected multisensory interactions in the perceived location of visual and auditory sources in the Causal Inference model



Here, the bimodal stimulation, $p(C)$, may be specified as $p(C = 1)$ and $p(C = 2)$, where $p(C = 1)$ stands for the probability that both visual and auditory events have the same underlying external cause. The value $p(C = 2)$ denotes the probability that both events have two separate external causes. Here, both causal structures are mutually exclusive and therefore must add up to one, namely,

$$p(C = 1) = 1 - p(C = 2). \quad (12)$$

Once the probability of the causal structure is established, one must model the percept associated to each causal structure. Here, the sensory estimate will be given by the MLE model, if a common cause is inferred ($p(C = 1)$):

$$\hat{s}_A|C = 1 = \hat{s}_V|C = 1 = \frac{x_A/\sigma_a^2 + x_V/\sigma_V^2 + p_{AV}/\sigma_p^2}{1/\sigma_A^2 + 1/\sigma_V^2 + 1/\sigma_{pAV}^2}, \quad (13)$$

where p_{AV} represents the combined prior of x_A, x_V . If two causes are inferred, then the sensory estimate is given by the no-interaction model:

$$p(\hat{s}_A|C = 2) = p(x_A|\hat{s}_A) p(A). \quad (14)$$

In the third and final step, the estimates from each underlying causality must be combined to predict the observed probability functions. So far, three approaches have been tested for audiovisual localisation. In one approach, it was hypothesised that subjects would simply select the most likely estimate. They would form the estimate associated to the most likely causal structure, that is,

$$\hat{s}_{A|x_V, x_A} = \begin{cases} \hat{s}_{A|C=1} & \text{if } p(C = 1) > 0.5 \\ \hat{s}_{A|C=2} & \text{if } p(C = 1) < 0.5 \end{cases}. \quad (15)$$

In an alternative strategy, the observed estimates would correspond to the linear weighted sum of the two independent estimates, obtained for each underlying possible causality

$$\hat{s}_{A|x_V, x_A} = p(C = 1 | x_V, x_A) \hat{s}_{A,C=1} + p(C = 2 | x_V, x_A) \hat{s}_{A,C=2}. \quad (16)$$

A final strategy would be that subjects alternate responses, according to each causal structure in a rate that matches the probability of that causal structure itself,

$$\hat{s}_A = \begin{cases} \hat{s}_{A,C=1} & \text{if } p(C = 1 | x_V, x_A) > \zeta \\ \hat{s}_{A,C=2} & \text{if } p(C = 1 | x_V, x_A) < \zeta \end{cases}, \quad (17)$$

where ζ is sampled randomly from a uniform distribution [0:1]. In practice, the observed response distributions are composed of two Gaussian distributions, each

with the relative size of each causal probability. There has been increasing evidence to support the probability matching strategy outlined in (17) (Wozny et al. 2010; Mendonça et al. 2016). This is counterintuitive, because it is not the optimal response strategy. A linear weighted sum would be the strategy with the least cost. Responding in such a way that matches the causal probability itself may mean one of two things:

1. Sometimes subjects might infer one cause and other times infer two causes. They may distribute their responses according to this pattern.
2. Alternatively, subjects might always access both possible percepts ($\hat{s}_A|C = 1$ and $\hat{s}_A|C = 2$). Then they may distribute the answers across trials according to a fluctuating internal criterion that accounts for the likelihood of a common source.

When the *causal inference model* was proposed (Körding et al. 2007; Sato et al. 2007; Beierholm et al. 2008; Wozny et al. 2010), it was solved using a generative model. Equations 13 and 14 were solved by assuming that all elements were Gaussian. To test the causal inference model, experiments on audiovisual localisation were carried out. In these tests, subjects indicated where they perceived the visual and auditory sources under unimodal and bimodal conditions. When testing the model, several parameter values were considered unknown. The causal inference model was fitted to experimental data, to find which model values fit the data best. The unknown parameters were the probability of common cause, $p(C = 1)$; the standard deviation of the visual representation, σ_V ; auditory representation, σ_A ; and of the prior (σ_p). The prior itself can be assumed to be an unknown parameter too. The causal inference model was found to be superior to other models of multisensory interaction in all of the studies mentioned above.

Causal Inference Without a Generative Model

Mendonça et al. (2016) proposed a solution that avoided the use of a generative model. Other multisensory integration models, like the MLE, have bypassed generative models by obtaining these parameters directly from experimental data—e.g., Ernst and Banks (2002); Alais and Burr (2004); Binda et al. (2007); Mendonça et al. (2011). In the model by Mendonça et al. (2016), data from unisensory conditions is collected and used to predict perception in multisensory conditions. In this case, the independent auditory localisation estimate, \hat{s}_A , can be observed directly by computing the mean of the experimentally obtained localisation responses (s_A). The standard deviation of the auditory localisation estimate, σ_A , is obtained as the standard deviation of that mean. The obtention of the internal estimates also spares from having to calculate the priors, $p(A)$ and $p(V)$. Any prior should affect the formation of the internal localisation estimate, \hat{s}_i , from the external source localisation, x_i —see (3), (14). If the estimate \hat{s}_i is observed directly, no additional biases are expected when the additional sensory cue is added to calculate the final multisensory estimate. Therefore, in this version of the causal inference model, the conditional estimates can be obtained as follows.

$$\hat{s}_{A,C=1} = \hat{s}_{V,C=1} = \frac{s_A/\sigma_A^2 + s_V/\sigma_V^2}{1/\sigma_A^2 + 1/\sigma_V^2}, \quad (18)$$

$$\hat{s}_{A,C=2} = s_A ; \hat{s}_{V,C=2} = s_V. \quad (19)$$

As learned from (13), when a common cause is observed, the visual and auditory estimates will be identical. Therefore the following equation is used in this model to observe the probability of common cause $p(C = 1)$, that is,

$$\begin{aligned} p(C = 1 | x_V, x_A) &= p(\hat{s}_A = \hat{s}_V) \\ &\text{and} \\ p(C = 2 | x_V, x_A) &= 1 - p(\hat{s}_A = \hat{s}_V). \end{aligned} \quad (20)$$

This model has been shown to simulate the perceived position of auditory image in distance accurately, and it performs better than the traditional models of sensory dominance, including the MLE and no-interaction models (Mendonça et al. 2016). Figure 8 shows the predictions of this model against data of sound distance estimations.

3.3 Sensory Weights

The computation of the sensory weights is informative on its own. The weight attributed to each sensory cue denotes how reliable each sensory cue is. The sensory weights may also provide an indirect indication of how accurately the signal is represented in the brain in a given context and task. In the case of spatial visual and auditory representations, it is known that the visual representations are more reliable than the auditory ones. In most of the models mentioned above, this is the main reason why visual cues typically dominate over auditory space judgements.

According to the MLE theory, the sensory weights can be calculated following (8). However, that theory has been shown to be incomplete. Mendonça et al. (2016) proposed a new way of calculating sensory weights that accounts for causal inference. It calculates separately the weights of the sensory cues in each underlying causal structure and sums them up accounting for the probability of that causal structure,

$$\begin{cases} w_A = w_{A,C=2} p(C = 2|x_A, x_V) + w_{A,C=1} p(C = 1|x_A, x_V), \\ w_V = w_{V,C=2} p(C = 2|x_A, x_V) + w_{V,C=1} p(C = 1|x_A, x_V), \\ w_A + w_V = 1. \end{cases} \quad (21)$$

A depiction of sensory weights calculated in this manner is presented in Fig. 9. This figure describes the weights of the visual and auditory cues on the visual and auditory distance estimates. Firstly, it can be observed that the sensory weights follow very different patterns for the visual and auditory distance estimates. This may indicate

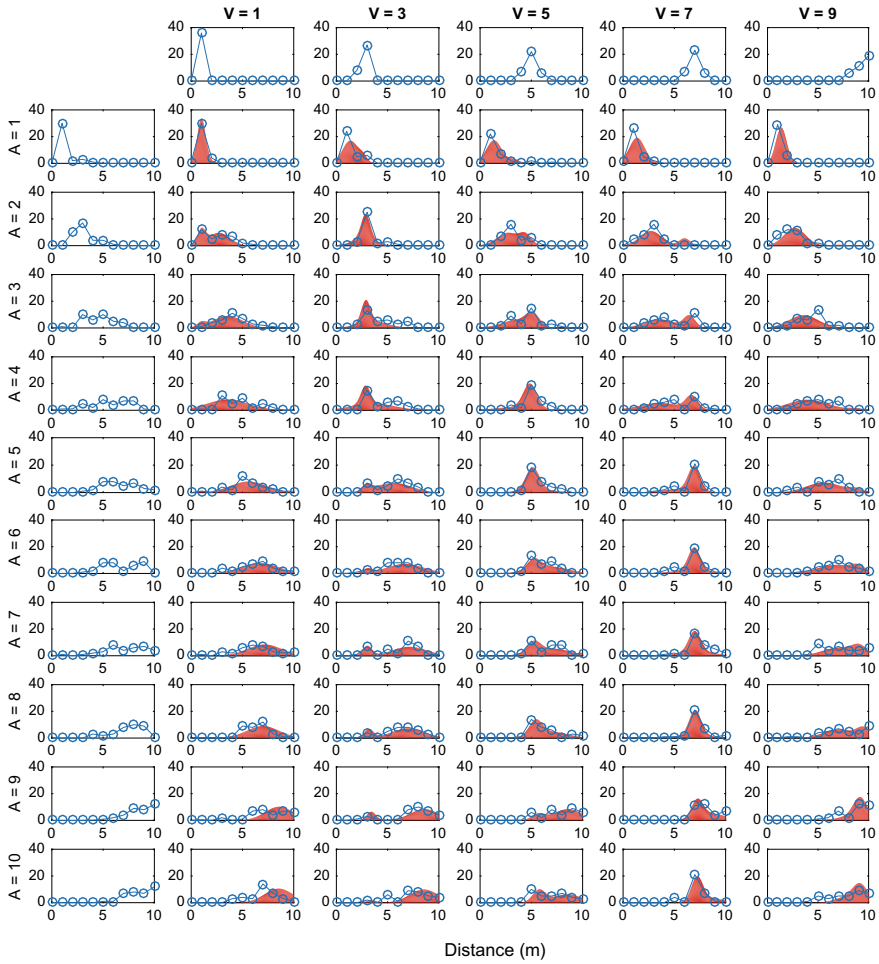


Fig. 8 The predictions of the causal inference without generative model against sound localisation data in distance perception—adopted from Mendonça et al. (2016). The top row and left column display the distribution of localisation estimates in the visual and auditory unimodal conditions, respectively. The blue line corresponds to the response distributions obtained in a distance estimation experiment. The red areas correspond to the predicted distributions from the causal inference model without a generative self-fitting approach

that, despite multisensory interactions, both estimates are processed separately. It can also be seen that the visual cue only achieves a higher weight over the auditory estimate if the stimuli are co-localised, or exist within a narrow window of space. This area, where the concurrent cues gain higher weights, may be conceptualised as the multisensory interaction window. The audiovisual interaction windows obtained through this method for sound distance localisation can be seen in Fig. 3.

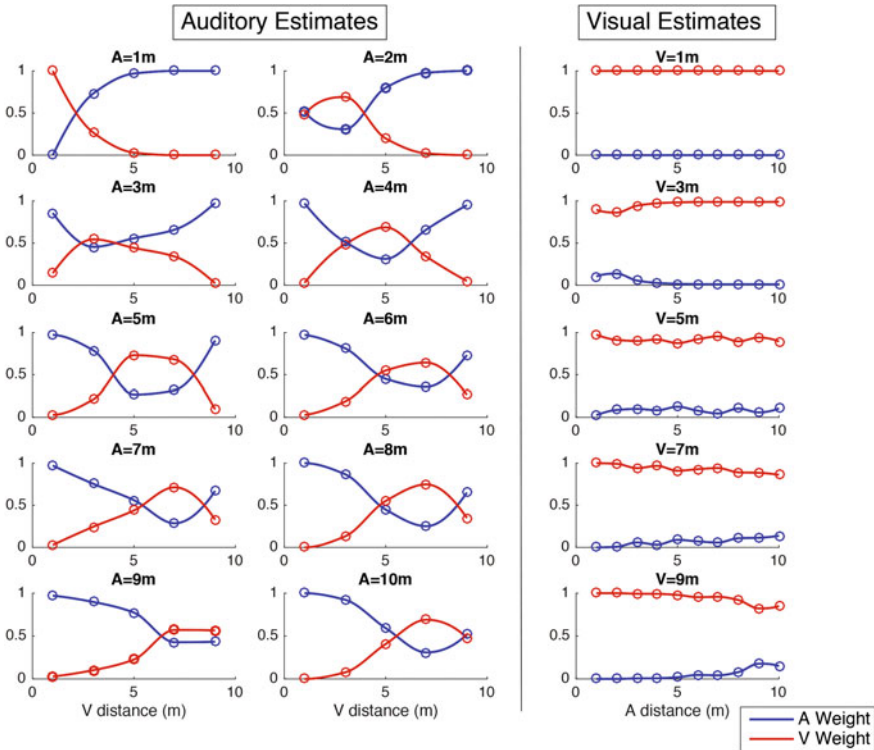


Fig. 9 Weights of the visual and auditory cue on the visual and auditory distance estimates, as obtained by Eq. 21. The area where the weight of the visual cue over the auditory estimate rises can be determined as the spatial window of multisensory interaction

4 Modelling Sound Localisation After Audiovisual Stimulation

As discussed earlier in the chapter, after exposure to spatially discrepant audiovisual stimulation, the auditory spatial map becomes shifted in the direction of the shift observed during the audiovisual exposure. This effect is known as the *ventriloquism aftereffect*. There have been two different approaches to modelling sound localisation that account for previous audiovisual experience. The first approach applies the causal inference model described above, and the second one analyses preceding sensory experience and sequential effects.

4.1 Causal Inference

Causal inference has been applied to simulate the ventriloquism aftereffect by Sato et al. (2007). This model was also applied to the ventriloquism effect itself. The causal inference model by Sato et al. (2007) is different from the models presented above by further specifying that the probability of common cause, $p(C)$, is affected by spatial and temporal parameters. To model the ventriloquism aftereffect, the authors applied a causal inference concept with a generative model that updates the spatial percepts $p(s_A|x_A)$ and $p(s_V|x_V)$ on a trial-by-trial basis including two additional free parameters μ_A and μ_V . Note that this model did not include the priors $p(A)$, $p(V)$, or $p(AV)$. When the spatial estimate $p(s_A|x_A)$ follows a normal distribution, the probability distribution function can be computed following a Gaussian distribution,

$$p(s_A|x_A) = \frac{1}{\sqrt{2\pi} \sigma_A^2} e^{-\frac{(s_A - x_A - \mu_A)^2}{2\sigma_A^2}}, \quad (22)$$

where σ_A^2 is the standard deviation of $p(s_A|x_A)$. A similar equation is used to compute the perceived visual location. In this case, the generative model explicitly starts with an unbiased aftereffect value, $\mu_A = 0$. The model, therefore, assumes an unbiased observer. After each audiovisual stimulation, the value of the parameter μ_A is updated as follows,

$$\mu_A \rightarrow (1 - \alpha_A)\mu_A + \alpha_A(s_A - \hat{s}_A), \quad (23)$$

where α_A is the magnitude of the adaptation after each audiovisual stimulus. It is observed from the biases in response to previous stimuli. In simulations, this model was found to follow the results of the ventriloquism aftereffect obtained by Recanzone (1998). As noted by the authors, this model is promising, but also presents some limitations. One main limitation is that this model fails to predict any effect or aftereffect when visual and auditory sources are too far apart. The causal inference model requires the sources to be close enough so that they can be interpreted as having one single external cause. When they are too far apart, they are perceived as two separate events, and therefore there should be no aftereffect. In contrast to the model performance, it has been shown that the ventriloquism aftereffect occurs even when sources are separated by more than 20° .

4.2 Inverse Model of Recent Experience

Mendonça et al. (2015) took an exploratory approach to find which audiovisual experience parameters primarily influence the auditory spatial shift. Varying sequences of audiovisual events, which could be congruent or discrepant in space, were presented to subjects. Their estimates of light and sound source localisation were used to model

Table 1 Models tested in each **P** matrix. The parameter $AICW_i$ denotes the relative probability of each model best describing the auditory shifts, according to Akaike’s Information Criterion

Model	$AICW_i$
(A) $SHA = SQ_1 w_1 + SQ_2 w_2 + SQ_3 w_3 + SQ_4 w_4 + SQ_5 w_5$	0.143
(B) $SHA = SQ_1 w_1 + SQ_2 w_2 + SQ_3 w_3 + SQ_4 w_4 + SQ_5 w_5 + D w_D$	0.004
(C) $SHA = SQ_1 w_1 + SQ_2 w_2 + SQ_3 w_3 + SQ_4 w_4 + SQ_5 w_5 + MaxD w_{MaxD}$	0.004
(D) $SQ_5 w_5 + D w_D$	0.655
(E) $SQ_5 w_5 + MaxD w_{MaxD}$	0.000
(F) $SQ_5 w_5 + D w_D + MaxD w_{MaxD}$	0.194

how recent experience affected the auditory localisation shifts. It was hypothesised that the auditory shift could be predicted by

$$SH_A = \sum_i^n w_{P_i} P_i, \tag{24}$$

where SH_A is the auditory localisation shift. It is observed as the difference between a physical sound source position and its perceived location. The variable P_i stands for each tested parameter, and w_{P_i} is the weight of this parameter. Several equations were conceived, with a different number of parameters, to test which one would fit the data best. The most successful models are presented in Table 1. One model added each of the last five audiovisual stimuli (SQ_1, \dots, SQ_5) as individual parameters. Others included not only each of the previous trials, but also the total number of discrepant trials experienced in that period, D ; the maximum number of discrepant trials in a row, $MaxD$; and all of the above. It was also tested whether the last trial alone, SQ_5 , and the last trial combined with D and $MaxD$ would fit the data better.

The weights were obtained through a least-squares fit. For each model, a matrix **P** was created with each stimulation type as a column vector and each tested parameter as a row vector. Therefore, **P** had as many columns as the number of parameters P_i and as many rows as all the possible combinations of parameter values. For example, to test Model A from Table 1, the matrix **P** was configured as follows:

$$\mathbf{P} = \begin{bmatrix} SQ_{1,0} & SQ_{2,0} & \dots & SQ_{5,0} \\ SQ_{1,1} & SQ_{2,0} & \dots & SQ_{5,0} \\ \vdots & \vdots & \ddots & \vdots \\ SQ_{1,1} & SQ_{2,1} & \dots & SQ_{5,1} \end{bmatrix}.$$

The subsequent auditory localisation shifts were added to another matrix **SH**, with a single column, where each row value contained the average auditory localisation shift observed after the stimuli described in the corresponding row of **P**. The weights of each parameter **W** are the multiplier of **P** with the product **SH**:

$$\mathbf{SH} = \mathbf{P} * \mathbf{W}. \quad (25)$$

The best fitting values of \mathbf{W} are given by the pseudo inverse matrix, obtained by

$$\mathbf{W} = (\mathbf{P}^T * \mathbf{P})^{-1} * \mathbf{P}^T * \mathbf{SH}. \quad (26)$$

All models presented in Table 1 were subsequently tested by fitting the model to the data in a linear regression test. Since models with more parameters benefit from having more degrees of freedom to fit the data, those models will always have an advantage. They tend to overfit. Therefore, a correction has to be applied. A modified Akaike's information criterion (AIC_C) was used (Akaike 1974), allowing for the number of parameters to be accounted for. Akaike's model weights ($AICW_i$) are presented in Table 1. According to this criterion, the higher the weight, the better is the model. It was found that Model D performed best using this criterion. Considering only the last audiovisual trial and the overall number of discrepant trials in recent sensory experience led to the best prediction, namely,

$$SH_A = SQ_5 W_{SQ5} + DW_D, \quad (27)$$

where $W_{SQ5} = 0.88$ and $W_D = 3.42$. Therefore, the number of discrepant trials in the recent sensory experience stands out as the highest weighted parameter in predicting auditory space calibration following audiovisual experience.

5 Discussion, Future Directions and Conclusion

This book chapter presented an overview of psychophysical phenomena and psychophysical models in sound localisation during and after audiovisual stimulation. This topic has been the subject of research for several decades, but the field is more dynamic than ever, with newer models being proposed more frequently over the last decade.

Regarding psychophysical effects, it has been established that visual information has the potential to impact the localisation of sound sources in azimuth, regardless of how far apart the light and sound sources are in space. When presented within close proximity, the sound source may be perceived as co-localised with the visual source. If not, the sound source will still be perceived as closer to the visual source than it was presented. In sound localisation in elevation, there is a surprising scarcity of research. Current findings point to an influence of visual cues, which is similar to the one found for horizontal localisation, but not higher for elevation. This is surprising because most models attribute higher weight to cues with lower estimate variability. In elevation, binaural cues provide little information for sound localisation, and subjects must rely mostly on monaural cues. Therefore, there is greater inaccuracy and higher uncertainty in identifying sound source elevation, as compared

to azimuth. The effects of this uncertainty have never been carefully tested in regards to multisensory integration.

Concerning distance perception, vision remains more accurate than hearing. Nevertheless, there is a remarkable auditory robustness. Unlike for angular judgments, concurrent visual information is unable to affect the perceived sound source distance unless both sources are in close proximity to each other. Despite of these differences, the Bayesian causal inference model still performs best in describing those multisensory interactions.

Psychophysical models of sound localisation during audiovisual stimulation are not exclusive for localisation, and they can also be tested in other contexts of multisensory interaction. These models have evolved greatly over the last decades, and, interestingly, they are mostly evolving cumulatively. For the most part, the newest models include elements of the old models. The concept of visual dominance and ventriloquism evolved from a winner-takes-all view to primarily represent the strong impact of the visual cue over the auditory space estimate. The modality appropriateness hypothesis suggested for the first time that the impact of the visual cue was mostly related its reliability. The MLE model simulates multisensory interactions, and it accounts for cue reliability by explicitly giving a weight to each sensory cue. It is also the first model to assume an optimal observer, who maximises the model performance in presence of sensory noise. This model proved to be robust in describing human perception. However, it did not account for situations where cues were not integrated, where there was no mutual effect of the cues upon each other, or when this effect was small. The Bayesian causal inference model accounts for the multisensory integration predicted by the MLE and for the absence of interaction in the no-interaction model. It also predicts the intermediate cases as a combination of those two scenarios. Bayesian causal inference is the most commonly accepted model to describe audiovisual interactions in localisation. This model has also been applied to other contexts, not described here. For instance, a variation of this model has been used under the framework of Bayesian networks to model the process of the multisensory interactions in audiovisual localisation (Besson et al. 2010). There is also some neurophysiological evidence to support the idea that the brain behaves similar to the causal inference model (Kayser and Shams 2015).

Concerning the models, currently the main challenge is to find out which strategy is used to estimate sound localisation when the probability of common cause lies in between total-integration or no-interaction cases. Current data indicates that subjects simply alternate between perceiving one or the other—the integration or the no-interaction case—in a way that matches the underlying probability of each case. Consciously alternating responses may explain varying perception during the experiment. The two options should be differentiated. There are also reports of localisation at intermediate positions between those predicted by each modality. This outcome corresponds to a different mechanism of integrating the two modalities. Identifying exactly which mechanisms take place, and the level of automation versus conscious decision should be addressed in future research.

Another challenge is to model sound localisation after audiovisual exposure. Bayesian causal inference fails to provide an actual prediction of the magnitude

of the changes. This model can only be used to describe what happened after it has been observed by looking into the changes in localisation. It does not account for temporal adaptation or type of exposure. The inverse model of recent experience reveals that a weighted linear sum of the overall amount of discrepant audiovisual experience, combined with the type of experience observed in the very last audiovisual trial predict the amount of displacement accurately in subsequent sound localisation. However, this model is dependent on the data itself, like the Bayesian causal inference with a generative model and other self-fitting models. Self-fitting models leave model parameters unspecified, and they find the values for those parameters that will provide the best fit for the data. Even after applying correction factors like Akaike's criterion, the generalisation of the findings is limited, and it is impossible to ascertain how valid the model is. More research will be needed to validate these models and, ideally, new models should emerge that provide independent predictions.

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant No. 659114. Two anonymous reviewers have contributed with very constructive remarks.

References

- Akaike, H. 1974. A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, 215–222, Springer.
- Alais, D., and D. Burr. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* 14 (3): 257–262.
- Beierholm, U., L. Shams, W.J. Ma, and K. Koerding. 2008. Comparing Bayesian models for multisensory cue combination without mandatory integration. In *Advances in Neural Information Processing Systems*, 81–88.
- Bertelson, P. 1999. Ventriloquism: A case of crossmodal perceptual grouping. *Advances in Psychology* 129: 347–362.
- Bertelson, P., and M. Radeau. 1981. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Attention, Perception, & Psychophysics* 29 (6): 578–584.
- Besson, P., J. Richiardi, C. Bourdin, L. Bringoux, D.R. Mestre, and J.-L. Vercher. 2010. Bayesian networks and information theory for audio-visual perception modeling. *Biological Cybernetics* 103 (3): 213–226.
- Binda, P., A. Bruno, D.C. Burr, and M.C. Morrone. 2007. Fusion of visual and auditory stimuli during saccades: a Bayesian explanation for perisaccadic distortions. *Journal of Neuroscience* 27 (32): 8525–8532.
- Bresciani, J.-P., F. Dammeyer, and M.O. Ernst. 2006. Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision* 6 (5): 2–2.
- Choe, C.S., R.B. Welch, R.M. Gilford, and J.F. Juola. 1975. The “ventriloquist effect”: Visual dominance or response bias? *Attention, Perception, & Psychophysics* 18 (1): 55–60.
- Ernst, M.O., and M.S. Banks. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415 (6870): 429–433.
- Hartline, P.H., R.P. Vimal, A. King, D. Kurylo, and D. Northmore. 1995. Effects of eye position on auditory localization and neural representation of space in superior colliculus of cats. *Experimental Brain Research* 104 (3): 402–408.
- Held, R. 1955. Shifts in binaural localization after prolonged exposures to atypical combinations of stimuli. *The American Journal of Psychology* 68 (4): 526–548.

- Kaysers, C., and L. Shams. 2015. Multisensory causal inference in the brain. *PLoS Biology* 13 (2): e1002075.
- King, A.J. 2009. Visual influences on auditory spatial learning. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1515): 331–339.
- Körding, K.P., U. Beierholm, W.J. Ma, S. Quartz, J.B. Tenenbaum, and L. Shams. 2007. Causal inference in multisensory perception. *PLoS One* 2 (9): e943.
- Landy, M.S., L.T. Maloney, E.B. Johnston, and M. Young. 1995. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research* 35 (3): 389–412.
- Laurienti, P.J., R.A. Kraft, J.A. Maldjian, J.H. Burdette, and M.T. Wallace. 2004. Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research* 158 (4): 405–414.
- Mendonça, C., A. Escher, S. van de Par, and H. Colonius. 2015. Predicting auditory space calibration from recent multisensory experience. *Experimental Brain Research* 233 (7): 1983–1991.
- Mendonça, C., M. Hiipakka, S. van de Par, and H. Colonius. 2014. Adaptation to non-individualized spatial sound through audiovisual experience. In *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Audio Engineering Society.
- Mendonça, C., P. Mandelli, and V. Pulkki. 2016. Modeling the perception of audiovisual distance: Bayesian causal inference and other models. *PLoS One* 11 (12): e0165391.
- Mendonça, C., J.A. Santos, and J. López-Moliner. 2011. The benefit of multisensory integration with biological motion signals. *Experimental Brain Research* 213 (2–3): 185.
- Meredith, M.A., and B.E. Stein. 1986. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology* 56 (3): 640–662.
- Posner, M.I., M.J. Nissen, and R.M. Klein. 1976. Visual dominance: An information-processing account of its origins and significance. *Psychological Review* 83 (2): 157.
- Radeau, M., and P. Bertelson. 1974. The after-effects of ventriloquism. *The Quarterly Journal of Experimental Psychology* 26 (1): 63–71.
- Recanzone, G.H. 1998. Rapidly induced auditory plasticity: The ventriloquism aftereffect. *Proceedings of the National Academy of Sciences* 95 (3): 869–875.
- Roach, N.W., J. Heron, and P.V. McGraw. 2006. Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society of London B: Biological Sciences* 273 (1598): 2159–2168.
- Rock, I., and J. Victor. 1964. Vision and touch: An experimentally created conflict between the two senses. *Science* 143 (3606): 594–596.
- Sato, Y., T. Toyoizumi, and K. Aihara. 2007. Bayesian inference explains perception of unity and ventriloquism aftereffect: Identification of common sources of audiovisual stimuli. *Neural Computation* 19 (12): 3335–3355.
- Silva, C.C., C. Mendonça, S. Mouta, R. Silva, J.C. Campos, and J. Santos. 2013. Depth cues and perceived audiovisual synchrony of biological motion. *PLoS one* 8 (11): e80096.
- Slutsky, D.A., and G.H. Recanzone. 2001. Temporal and spatial dependency of the ventriloquism effect. *Neuroreport* 12 (1): 7–10.
- Van Eijk, R.L., A. Kohlrausch, J.F. Juola, and S. van de Par. 2008. Audiovisual synchrony and temporal order judgments: Effects of experimental method and stimulus type. *Attention, Perception, & Psychophysics* 70 (6): 955–968.
- Vroomen, J., P. Bertelson, and B. De Gelder. 2001. The ventriloquist effect does not depend on the direction of automatic visual attention. *Attention, Perception, & Psychophysics* 63 (4): 651–659.
- Wallace, M.T., L.K. Wilkinson, and B.E. Stein. 1996. Representation and integration of multiple sensory inputs in primate superior colliculus. *Journal of Neurophysiology* 76 (2): 1246–1266.
- Welch, R.B., and D.H. Warren. 1980. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin* 88 (3): 638.
- Werner, S., J. Liebetrau, and T. Sporer. 2013. Vertical sound source localization influenced by visual stimuli. *Signal Processing Research* 2 (2): 29–38.

- Wozny, D.R., U.R. Beierholm, and L. Shams. 2010. Probability matching as a computational strategy used in perception. *PLoS Computational Biology* 6 (8): e1000871.
- Wozny, D.R., and L. Shams. 2011. Recalibration of auditory space following milliseconds of cross-modal discrepancy. *Journal of Neuroscience* 31 (12): 4607–4612.
- Yuille, A.L., and H.H. Bulthoff. 1996. Bayesian decision theory and psychophysics. *Perception as Bayesian Inference*, 123.

Cross-Modal and Cognitive Processes in Sound Localization



M. Torben Pastore, Yi Zhou and William A. Yost

Abstract To perceptually situate a sound source in the context of its surrounding environment, a listener must integrate two spatial estimates, (1), the location, relative to the listener's head, of the auditory event associated with the sound-source and, (2), the location of the listener's head relative to the environment. This chapter introduces the general background of auditory localization as a multi-sensory process and reviews studies of cross-modal interactions with auditory localization for stationary/moving sound sources and listeners. Included are relevant results from recent experiments at Arizona State University's Spatial-Hearing and Auditory Computation and Neurophysiology Laboratories. Finally, a conceptual model of the integrated multisensory/multi-system processes is described.

1 Introduction

Sound-source localization is a part of the larger perceptual process wherein transduced sensation is analyzed to form an internal representation of the surrounding environment, including the listener's own position in it. The internal reference created by this process is often called a *spatial map*. For a review of spatial maps, see Stensola and Moser (2016). Localizing a sound source in relation to other perceived objects requires mapping the first-level auditory spatial estimate, which only relates sound-source position to the listener's head, into the context of the surrounding local environment.

Consider an attempt to localize a sound source without such context. Perceptually salient sound stimulation must be parsed into individual perceptual objects, perhaps in interaction with other sensory inputs such as vision. Having grouped a set of components of the sound stimulation into a specific auditory object to be localized, the listener must then extract auditory spatial cues by comparing the inputs at the two ears across frequency as well as amplitude and phase patterns across frequency. The

M. T. Pastore (✉) · Y. Zhou · W. A. Yost
College of Health Solutions, Arizona State University, Tempe, AZ 85287, USA
e-mail: m.torben.pastore@gmail.com

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_12

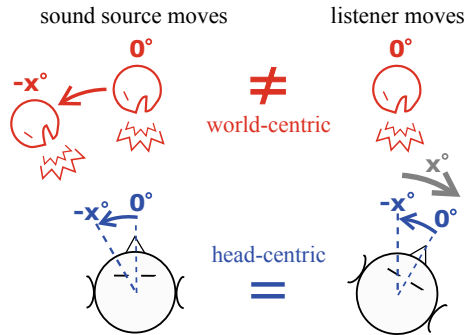


Fig. 1 A schematic illustration of the difference between head-centric and world-centric auditory localization. The actual sound source, located in the local environment, is shown in **red** with its angular displacement noted above in world-centric coordinates. In **blue**, the angular displacement of the sound source vis-à-vis the listener's head, that is, in head-centric coordinates, is shown. Positive values indicate clockwise displacement. In the **left panel**, the sound source moves from the midline to $-x^\circ$ in world-centric coordinates, resulting in a change in location relative to the listener's head from the midline (0°) to $-x^\circ$. In the right column, the sound source is stationary at 0° in world-centric coordinates, but the listener rotates the head by $+x^\circ$, as shown by the **gray arrow**, so the head-centric estimate of the sound-source location becomes $-x^\circ$, that is, the same as in the **left panel**. The listener must know the position of the head in relation to the local environment to localize the sound source in world-centric coordinates

result is some estimate of the location of the sound source relative to the listener's head. Without further information, the listener cannot utilize this perceptual output for action, because there is, so far, no internal representation of the space around the listener. Figure 1 illustrates this concept. Without information about the listener's head position, the dynamic auditory spatial cues are the same for a sound source that moves while the listener is stationary versus a stationary sound source while the listener moves ($-x^\circ$, printed in blue in Fig. 1). The two are therefore indistinguishable. Even with information about the listener's head position, the listener still only knows the location of the sound source relative to the head. To determine the location of the sound source relative to the surrounding environment, the listener must know the orientation of the head relative to the body and the local environment.

It is precisely because creating a perceived spatial map requires an estimate of one's location in that internally constructed context that the senses must rely on each other for reference, and that systems inputs—such as somatosensory, kinesthetic, muscular efferents, and proprioception—will necessarily interact with auditory spatial estimates at some level. Reduced to its simplest components, localizing a sound source in relation to the local environment requires mapping the estimate of the location of the sound source, relative to the listener's head, onto an internal representation of the local environment; this requires an internal representation of the listener's head position relative to the body and the surrounding environment. While this may seem obvious, the process by which this occurs is not. Many questions arise. For example, does mapping the auditory estimate into a spatial estimate of the local environment

occur at peripheral, midbrain, auditory cortex, or higher levels associated with cognitive processing—or all (some) of the above? What are the inputs to this process, and how are they combined and compared with each other? Does this combination occur according to a static rule or as a dynamic process that changes according to some set of internal and external factors, perhaps based on estimates of the reliability of the different inputs? What sort of auditory localization is possible when the internal estimate of the listener's location within the local environment is incomplete or the surrounding environment is perceptually inscrutable? Much remains to be done to address these types of questions.

This greater synthesis is likely to involve sensory, sensorimotor, and cognitive inputs. In other words, auditory localization is ultimately not merely a sensory task—it also engages non-sensory processes such as memory, attention, expectation, and motor signals. All of these questions lead inexorably to the conclusion that to fully understand spatial hearing, current inquiries must be expanded to include neural processes that occur *outside* the auditory system. Wallach (1938, 1939, 1940) was perhaps the first scientist in the modern era to enunciate and investigate these considerations. For this reason, a considerable portion of this chapter is devoted to the points he made in his seminal works on this subject. Most of the literature considered in this chapter attempts to extend findings from the laboratory toward the daily, real-world task of localizing sound sources as listeners and/or sound sources move.

The concluding section of this chapter describes the scope of this greater inquiry via a model that conceptually organizes the seemingly disparate investigations that have been reported in the literature. The model may then be used to identify future areas of study necessary to understanding auditory localization as a multi-systems/multisensory process.

2 General Review

2.1 *Theories of Sound-Source Localization Before the 20th Century*

The early study of sound-source localization in the mid-19th century was based almost entirely on assumptions regarding the use of other sensory systems or experience in using sound to locate sound sources in the actual world—see Boring (1942). The question of whether the mind is different to the body in kind or only in degree—Cartesian Mind-Body Dualism—was a major topic in science and philosophy. Several scholars argued that the mind represents properties of the external world through sensations. These sensations had attributes, such as quality, intensity, duration, and extension, and they could be used to form percepts that the mind could use to create an internal representation of the external world. Scholars debated the exact definitions and means of measuring sensations, attributes, and perception (the mind) for nearly a half century. During this time, several scholars addressed sound-source localization.

Originally, most argued that sound has no attributes of extension (size and shape) when it impinges on the eardrum, so listeners could not use sound, on its own, to locate a sound source. This, of course, flew in the face of what most could observe, namely, that listeners can indeed localize sound sources using their hearing. Empiricists like Wundt argued that sound-source location was mediated by other senses that could sense extension, for instance, vision, touch, and the vestibular sense—see Boring (1942). Early psychologists, for example, Berkeley (1709), argued that experience helped in sound-source localization—see Pierce (1901). For at least 25 years, scholars therefore believed that sound-source localization resulted from interactions with other sensory systems and/or experience. As the 19th century ended, it became more and more accepted that sound has attributes associated with spatial extension. The question of cross-modal sound-source-localization processes became an item of increasing interest. For instance, Boring (1942) posed the question, “*Can the organism discriminate the relative positions of sounds, and, if so, how?*” This approach, exemplified in the work of Rayleigh (1876) and Thompson (1878), moved the view of sound-source localization as a multi-system process to one based on the ability of the mind (brain) to exploit differences in the inputs to the two ears to compute cues that could be used to estimate a source’s location, entirely based on the sound it produces.

2.2 Auditory Input for Stationary Listeners

The auditory spatial cues are well described and documented in the literature, especially those required for azimuthal localization, since the late 19th century (Boring 1942; Mills 1972; Blauert 1997; Yost 2017a). The auditory spatial cues are commonly described in terms of the three spatial dimensions (azimuth, elevation, and distance/range)—these are discussed below.

Interaural Differences of Time and Intensity

At any given elevation, interaural differences of time (ITDs) and level (ILDs) serve as the primary cues for estimating the azimuthal location of a sound source. In normal soundfield listening conditions (i.e., excluding headphone listening) ITDs dominate the localization of low-frequency sounds ($\lesssim 1300$ Hz, e.g., see Mills 1960; Macpherson and Middlebrooks 2002). Note that listeners *are* sensitive to low-frequency ILDs over headphones, and demonstrate roughly the same sensitivity to ILDs at all frequencies within the range of hearing (Yost 1981). However, in a soundfield the magnitude of low-frequency ILDs is typically small due to diffraction of long wavelengths around the head. The magnitude of high-frequency ILDs is considerably larger, and fine-structure ITDs at high-frequencies are poorly encoded, if at all, so ILDs are the dominant cue for localizing high-frequency sounds. ILDs of a decibel or more, the ILD difference threshold, are generally measured for frequencies greater than 2000 Hz—see Goupell and Stakhovskaya (2018) (but compare Hartmann et al. 2016). For further details refer to Kuhn (1977, 1987). Envelope ITDs

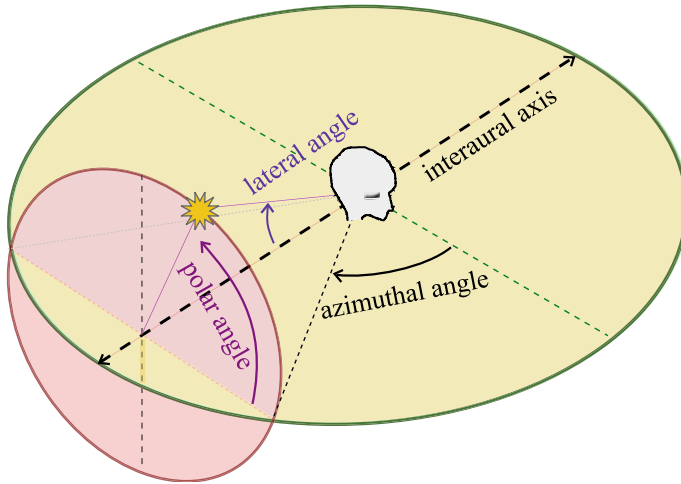


Fig. 2 The interaural double-pole coordinate system. The *interaural axis* is defined by that line which goes through both the listener’s ears. The sound source is illustrated as a large **yellow asterisk**. The *lateral angle* is the angle between the sound source and the interaural axis; it is thus a combination of azimuth and elevation. The *polar angle* is the angle between the sound source and the azimuth plane, along its sagittal plane, or “cone of confusion”. The *azimuthal angle*, which is complementary to the lateral angle, is the angle between the midline and the location where the sagittal plane (**in red**) meets the azimuthal plane (**in green**). Note that the elevation angle in the single-pole coordinate system (see, for example, Fig. 6), is not interchangeable with the polar angle. Thus, it is important to specify which system is being used. Figure adapted from Morimoto (2001)

can affect perceived lateral position when sounds are presented over headphones (e.g., Blauert 1997; Bernstein and Trahiotis 2011). However, recent studies presenting similar stimuli in a sound field (Macaulay et al. 2017; Yost 2017b) failed to find a similar effect for envelope ITDs—it may be that the presence of a strong ILD cue at these frequencies renders the envelope cue redundant.

Spectral-Shape Cues

Figure 2 shows a sagittal plane (in red) intersecting with the azimuthal plane (in green) in the *interpolar coordinate system* (also called the “two-pole” system, e.g., Letowski and Letowski 2011). At any angular location on the azimuthal plane, there is a locus of possible sound-source positions that generate the same interaural disparities, especially low-frequency ITDs. Note that the iso-contours for ILDs are more complex, and the pattern of ILDs across frequency may, in itself, be useful for specifying a unique sound-source location. These loci are the so-called *cones of confusion*, (see Wallach 1938; Woodworth and Schlosberg 1938) defined by the sagittal planes in the interaural polar coordinate system (see also Baumgartner et al. 2013). Spectral-shape cues created by the filtering of sound as it passes over the torso, head, and pinna on the way to the ear canal—the head-related transfer function (HRTF)—allow listeners to determine the location of a stimulus on that locus—i.e., its polar elevation, including whether it is in front of or behind the listener. Such HRTF cues

are most useful for broadband, high-frequency (>3000 Hz) sounds. For further information, see Morimoto and Aokata (1984), Middlebrooks et al. (1989), Makous and Middlebrooks (1990), Blauert (1997). HRTF cues can aid elevation estimations if listeners have prior information about a sound's spectrum (Wightman and Kistler 1997). It would seem possible that a listener might be able to use head movements to gain familiarity with the spectrum of a stimulus by averaging across "looks" during a head-turn, or simply noting the changes in the peaks and dips of the sound spectrum as the head moves, though the authors are unaware of any such study in the literature.

An exact description of the HRTF spectral features which are responsible for elevation judgments has not been agreed upon at this time. In light of the fact that listeners do not localize elevation well with generic, KEMAR¹ HRTFs, and yet can "learn" new HRTFs (e.g., Hofman et al. 1998; Zahorik et al. 2006; Carlile and Blackman 2014) it appears likely that different listeners use different features of their own, individual HRTFs (Wenzel et al. 1993). Therefore, it seems unlikely that there is any pattern of specific spectral features, such as dips versus peaks, that is used in the same way by all listeners (see Middlebrooks 1992; Langendijk and Bronkhorst 2002). For a review on modeling of localization along sagittal planes, see Baumgartner et al. (2013).

While spectral cues are often thought of as a monaural cue, the way the spectra of the two ears are combined or weighted against each other is still not fully understood. There is evidence that the spectral cues of the ear ipsilateral to the sound source are weighted increasingly as the distance of the sound source from the midline increases. Asymmetries of the head and ears may also provide an interaural spectral difference, though it appears subservient to "monaural" spectral cues (for more information see Searle 1973; Musicant and Butler 1984; Humanski and Butler 1988; Slattery and Middlebrooks 1994; Morimoto 2001; Van Wanrooij and Van Opstal 2004; Jin et al. 2004).

When stimuli do not have sufficient high-frequency information, the acuity of auditory localization in terms of azimuth is largely unaffected, but listeners' estimation of elevation is considerably degraded and front-back reversals occur quite often. Good and Gilkey (1996) tested localization in noise, thereby disrupting high-frequency spectral cues. They found that decreased signal-to-noise ratio negatively affected listeners' ability to distinguish front from back, had less impact on elevation accuracy and affected horizontal localization the least.

Interaction of Interaural Differences and Spectral Cues

When sound stimuli do not have high-frequency information, or the pinnae are occluded with ear molds to distort HRTF cues (e.g., Morimoto 2001), listeners often tend to localize sounds in those portions of the azimuth plane that intersect with the front and the back of the cone of confusion. This may result from learning that most salient sound sources lie roughly near the azimuth plane. Spectral cues appear to specify the location on the cone of confusion that corresponds to the location of the sound source (e.g., Morimoto and Aokata 1984; Best et al. 2011; Letowski and

¹KEMAR® is an often-used head-and-torso simulator—a so-called "dummy head".

Letowski 2011). Such a conception is subtly different to the idea that spectral cues encode elevation as it is specified in single-pole spherical coordinates, because the elevation is along the cone of confusion (i.e., on a sagittal plane), instead of being measured from the origin.

While Morimoto and Aokata (1984) and Makous and Middlebrooks (1990) have shown evidence that interaural differences and spectral cues may be estimated independently of each other, it is not clear how or at what point these two estimates are combined into a unified estimate of sound-source location. Also, the literature is somewhat mixed on whether interaural cues need to be correct for judgments of elevation to be accurate. In other words, if a listener cannot determine which cone of confusion the sound source is on, spectral cues may not be useful (for studies related to this question see Van Wanrooij and Van Opstal 2004; Morimoto 2001; Jin et al. 2004; Martin et al. 2004). It is also worth noting that the pattern of ILDs across frequency is not monotonic because of the acoustical bright spot that results from wave diffraction around the head (Macaulay et al. 2010). Therefore, the pattern of ILDs across frequency could conceivably also be used to specify where on a cone of confusion the sound source lies (e.g., Macpherson and Middlebrooks 2002). Section 4 discusses how listeners may, in the absence of spectral cues, use head movements to specify where on a cone of confusion a sound source lies.

Distance and Range

Distance cues seem to be almost completely based on listener expectations, and therefore require knowledge not only of how a sound source at a given distance relates to the head, but also of learned changes in the quality of a sound as it moves further away from, or closer to, a listener. There are several correlations between the distance of a sound source and its acoustical qualities that can be learned. If a sound source is in the near field (less than ≈ 0.3 m from the listener, depending on frequency), atypical ILDs result from the non-linear propagation of the sound around the head—this could offer a cue for judging distance (e.g., Brungart et al. 1999). For sound sources not in the near field, there are several other cues. Sounds from sources at large distances can be affected by the atmosphere, which acts as a low-pass filter, thereby providing a possible spectral cue for relative distance estimations that likely requires experience and expectation on the part of the listener (Kolarik et al. 2016). Sound intensity decreases with distance according to the inverse-square law—with expectation/memory this cue could also be exploited. In reverberant spaces, the direct-to-reverberant energy ratio decreases with increased distance, and provides a cue for judging relative distance (Zahorik 2002; Bronkhorst and Houtgast 1999). Note that this cue also relies on some expectation for the acoustics of the space. Auditory motion parallax may, in some cases, provide a cue for discerning relative sound-source distance (Genzel et al. 2018) and is discussed further below. See Kolarik et al. (2016) for a general review on auditory distance perception.

A Case for Multimodal Cues in Auditory Localization

The auditory spatial cues described above (excepting distance cues) are primarily head-centric cues. Expectation and a priori information—analyses of acoustic cues

that are based on experience—can provide indirect information to improve sound-source location. Wallach (1940) appears to have been the first to point out that the auditory spatial cues cannot, by themselves, specify the location of a sound source in the context of the local environment. Wallach (1940) demonstrated that “two sets of sensory data enter into the perceptual process of localization, (1), the changing interaural cues and, (2), the data representing the changing position of the head”—see Sect. 4 for further discussion.

While the spatial cues for sound-source localization (see above) have been well-researched for nearly 150 years, much less is known about how the cues used to estimate head position relate to sound-source localization. The literature is clear that vision is a vital cue for determining head position (Wallach 1940; Yost et al. 2015; Van Opstal 2016). The literature also suggests that additional auditory cues and/or vestibular, somatosensory, kinesthetic, proprioceptive, and neuro-motor control systems could also provide head-position information. Experience, coupled with memory as it manifests itself in spatial maps, might also provide head position information. For an exploration of some of the complexities inherent to this issue, see Buzsáki and Llinás (2017). Estimates of head (and body) position are therefore likely to be the product of a combination of cues and estimates arising from a wide range of sensory and systems inputs. The dynamic weighting of these head-position cues in determining head position, and how this weighting interacts with sound-source localization, is currently not well understood.

There is, however, a relatively rich literature on the integration of different spatial cues, related to other aspects of sound-source localization, that might also account for the integration of auditory spatial cues and head-position cues for world-centric sound-source localization. The next two sections consider evidence for sound-source localization as a multisensory/multi-systems process. Section 3 considers experiments probing audio-visual interactions under conditions where listeners and sound sources are stationary, and Sect. 4 considers investigations in which listeners and/or sound sources move.

3 Examples of Sound-Source Localization as a Multisensory Process—Localization with Stationary Listeners and Stationary Sound Sources

A great deal of study has been devoted to visual capture, in which visual stimuli affect the perceived sound-source locations. Vision is clearly an important sensory input for determining the location of the listener (body and head) with relation to the surrounding environment. Vision can perceptually situate a head-related auditory estimate of sound-source location into the spatial context of the surrounding environment. As such, interactions between audition and vision can be thought of as evidence for Wallach’s 1939/1940 insight before head movement is even considered.

When visual and auditory signals are both perceptually attributed to the same source, vision improves the accuracy of sound localization. Vision often plays a dominant role in spatial judgment. Spatial visual cues can override the spatial information of a sound, causing errors in sound localization. Commonly known as the *ventriloquism effect* or visual capture, the auditory event is localized to a seen source, even though the sound source is positioned at a different location (Howard and Templeton 1966).

A bias towards vision-centered experiments has meant that most of what is known about audio-visual interactions comes from localization results in the horizontal frontal field. Limited evidence, however, reveals that vision can also enhance auditory-distance estimation (Anderson et al. 2014). The role of vision (eyes open vs. closed) is more limited in vertical localization (Shelton and Searle 1980), though vision does appear important to the calibration of vertical localization—see, for example, Zwiers et al. (2001). The horizontal and vertical difference is likely a result of the different roles of eye and head movements in gaze orientation. Recently, Solomon et al. (2017) showed that eye movements preferentially exploit the horizontal span of the visual field. Head movements then shift this horizontal span up and down. Nevertheless, the vertical gaze of a listener can have a strong effect on perceived auditory elevation, as discussed in Sect. 3.2 below.

While vision is arguably involved in most everyday listening experiences, the common bias of vision over audition is not simply a result of relatively poor auditory spatial acuity. Indeed, when adequate localization cues (i.e. both ITDs and ILDs) are available across a sufficient range of frequencies, spatial hearing is remarkably accurate (Dorman et al. 2016; Yost 2016). The just-noticeable change in horizontal angular displacement, the minimum audible angle, can be as small as $1\text{--}2^\circ$ for sound sources near midline (Mills 1972; Hartmann and Rakerd 1989). Nevertheless, most of us rely primarily on vision when localizing objects around us, whether the objects make sound or not. This is probably because the auditory spatial estimate, on its own, only specifies the position of the sound source relative to the listener's head (Yost et al. 2015) whereas the spatiotopic encoding of vision is inherently world-centric.

Over the past decades, digital technology has greatly advanced the sophistication and automation of stimulus delivery and experimental procedures, helping to uncover, (1), the structural properties of auditory and visual stimuli that are conducive to cross-modal interactions and, (2), the cognitive factors (e.g., attention, expectation and experience) that affect listeners' assumptions and awareness of the origin and cause of the multisensory inputs (Radeau and Bertelson 1977; Welch and Warren 1980). The following sections summarize empirical evidence that addresses how active vision affects auditory localization performance via frame-of-reference and perceived target position. Also, how major differences between visual and auditory spatial mechanisms may affect estimates of the center, width, and front-back location of a perceived sound source is discussed. Further, the general framework of spatial audiovisual studies is dealt with and future directions for research relevant to real-life activities are discussed.

Numerous studies have investigated how vision affects sound-source localization, mostly in the horizontal plane. The general empirical findings related to several

hypotheses of vision's role are broadly summarized below. These hypotheses are not mutually exclusive and are evolving concepts.

- *The frame-of-reference hypothesis* Sound localization is more accurate when a listener can acquire, through free or voluntary eye movements, knowledge of the spatial layout of a lighted environment (Thurlow and Kerr 1970; Warren 1970; Platt and Warren 1972; Shelton and Searle 1980).
- *The visual-dominance hypothesis* Vision is a dominant sense in spatial tasks due to its superior spatial acuity. Vision can bias the perceived direction of a source of sound towards the direction of a visual cue (Jackson 1953; Choe et al. 1975; Bertelson and Radeau 1981).
- *The cue-reliability hypothesis* The reliability of estimates for each modality determines which sense dominates perception before they are combined. Reducing the saliency of visual cues weakens visual dominance (Battaglia et al. 2003; Alais and Burr 2004; Ernst and Bühlhoff 2004).

3.1 Relevance of the Frame of Reference

Boring (1926) suggested that listeners effectively map the perceived location of sound sources onto a spatial reference provided by vision. This hypothesis predicts that listeners will localize sound sources more accurately when their eyes are open, even if they cannot see the sound source—this is called visual facilitation. Warren (1970) demonstrated visual facilitation by hiding the spatial layout of a room and the loudspeakers within using a khaki cloth, so that the cloth alone constituted the “textured” environment for the task. Subjects hand-pointed to the perceived direction of a pulse-train auditory stimulus. The visual conditions were factorial combinations of eye open/closed, environment light/dark and vision free/fixated. Analyses compared the response error and response variability scores among various visual conditions.

Their results showed that active visual sensing of the physical layout of the environment, and objects in it, enhanced the acuity of listeners' auditory localization. On their own, free vision, a lit environment, or simply having the eyes open did not result in visual facilitation. The most favorable condition for visual facilitation was a combination of a lighted environment with free, target-directed eye movement. Performance under this condition was better than the lighted condition with a fixed gaze and the unlit condition with free eye movement. Warren argued that eye movement per se does not improve the accuracy of auditory localization, but that an illuminated visual environment allows better visual-motor (eye-hand) coordination by providing a spatial reference to guide action.

Shelton and Searle (1980) tested how vision affects the absolute identification of a sound-source position in a sound field. Half the subjects wore goggles painted over in black while the other half wore clear goggles. In all conditions, both sets of listeners could see the loudspeaker positions before and between testing sessions, so vision (together with memory for those listeners wearing the blacked-out goggles) could provide estimates of both the frame of reference and the target-source location. No

instructions were given to tell listeners where they could look. Listeners' auditory localization benefited most from vision with sound sources located in the frontal field along the horizontal axis. Vision also improved localization for sound sources located behind listeners and to their sides, but the improvement was far less than for the frontal horizontal span. However, there was no significant benefit to localization acuity along the vertical axis of the frontal field. These early data demonstrate that the limitation of human vision to the frontal field may have significant consequences on how auditory localization interacts with the knowledge of the frame-of-reference and target locations acquired through vision.

3.2 Relevance of Visual Target Cues

Over the past decades, multisensory research has provided a broad understanding of the spatial and temporal features of sensory stimuli that are conducive to cross-modal bias. The general conclusion is that visual bias is greater when sound and light stimuli come from sources positioned close to each other and/or are presented at the same time (Jackson 1953; Pick et al. 1969; Thurlow and Jack 1973; Choe et al. 1975; Jones and Kabanoff 1975; Slutsky and Recanzone 2001). This suggests that multisensory processing follows Gestalt perceptual grouping principles—that is, spatial and temporal proximity enhance fusion between audition and vision in establishing a unitary percept. Attention appears to play a limited role in the ventriloquist effect (Bertelson et al. 2000), suggesting that audio-visual interactions may occur at early sensory stages. Studies also show that perceptual fusion between auditory and visual events is not a necessary factor for visual bias. Partial or incomplete visual capture can occur even when the auditory and visual stimuli are not perceptually fused together (Welch and Warren 1980; Bertelson and Radeau 1981; Hairston et al. 2003; Wallace et al. 2004; Kording et al. 2007). Some degree of visual capture can also occur for asynchronously presented auditory and visual stimuli (Jack and Thurlow 1973; Thurlow and Jack 1973; Radeau and Bertelson 1974; Shelton and Searle 1980; Radeau and Bertelson 1987; Recanzone 2009). However, the strength of visual bias does decrease as the spatial and temporal separation between auditory and visual spatial estimates increases. Reviews include Welch and Warren (1980), Stein and Meredith (1990) and King (2009).

While the majority of audio-visual studies have emphasized the spatial and temporal conditions underlying multisensory interactions, separate lines of work reveal that the reliability of estimates (the inverse of the variance) for each modality determines which sense dominates the fused percept. This suggests that the dominant role of visual spatial information is scalable. Indeed, results have shown that reducing the saliency of visual cues by blurring or adding corruptive noise can weaken or even reverse visual capture (Ernst and Banks 2002; Battaglia et al. 2003; Alais and Burr 2004). These empirical results have been well described in a Bayesian framework, which establishes the relationship between the stimulus, S , and response, R . See Mendonça (2020), in this volume, and further, Sivia and Skilling (2006) for a review of Bayesian analysis.

The general principle of Bayesian estimates can be expressed in terms of the relationship between two conditional probabilities of stimulus and response, that is,

$$p(S|R)p(R) = p(R|S)p(S) \quad (1)$$

where $p(S|R)$ is the posterior probability, $p(R)$ is the marginal likelihood, $p(R|S)$ is the likelihood and $p(S)$ is the prior probability. With the assumption that the distribution of neural responses is constant and stable, the equation can be expressed as the proportionality

$$p(S|R) \propto p(R|S)p(S). \quad (2)$$

Equation (2) is the foundation of Bayesian-Inference theory. It states that the internal reconstruction of an event (the posterior probability) is the result of the likelihood estimate of whether this event leads to a neural response and an estimate of the stimulus distribution (the prior probability).

In the Bayesian model of audio-visual localization, it is assumed that auditory, A , and visual, V , cues are independently processed, $p(R_{AV}|S) = p(R_A|S) p(R_V|S)$. The modality-specific, neural representations, the likelihood estimates $p(R_A|S)$ and $p(R_V|S)$, typically consist of a one-to-one mapping of the auditory and visual cues associated with the position variable, in the form of a Gaussian function, $\mathcal{N}(\mu_A, \sigma_A^2)$ and $\mathcal{N}(\mu_V, \sigma_V^2)$, where $1/\sigma_V^2$ and $1/\sigma_A^2$ describe the reliability of neural estimates of the visual and auditory spatial cues, respectively. A large σ signals a greater uncertainty in the neural estimate with weak responses from many spatial channels. A small σ signals a reliable neural estimate with strong responses from selected spatial channels.

One may, for the moment, assume that the combined A and V cues lead to a fused percept (e.g., the ventriloquist effect) and that the prior distribution is flat ($p(S) = 1$). Given these assumptions, Battaglia et al. (2003) and Alais and Burr (2004) showed that the combined multisensory estimate (i.e., the mean of the posterior estimation) is equal to the weighted sum of the individual, unitary A and V estimates,

$$\mu_{AV} = \sigma_{AV}^2 \left(\frac{1}{\sigma_V^2} \mu_V + \frac{1}{\sigma_A^2} \mu_A \right). \quad (3)$$

The term σ_{AV}^2 describes the variance of the combined estimate, which is always smaller than the variances of the unisensory estimates, σ_A^2 and σ_V^2 , as follows,

$$\sigma_{AV}^2 = \left(\frac{1}{\sigma_V^2} + \frac{1}{\sigma_A^2} \right)^{-1} = \frac{\sigma_A^2 \sigma_V^2}{\sigma_A^2 + \sigma_V^2} \leq \min(\sigma_A^2, \sigma_V^2). \quad (4)$$

When experimentally manipulating σ_A^2 and σ_V^2 it is important to carefully consider fundamental differences in the peripheral mechanisms of vision and audition. The visual peripheral system is spatiotopically organized—thus, it encodes space

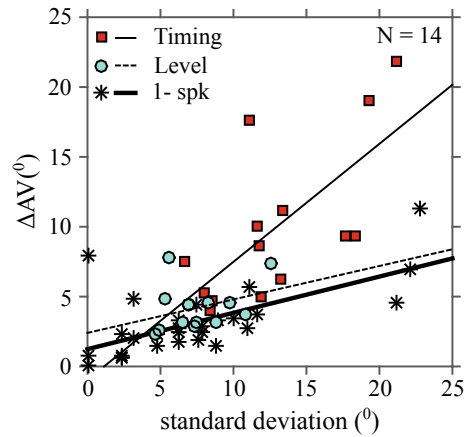
directly. The receptive fields of ganglion cells cover different regions of space that are mapped onto the retina, and the visual system retains this mapping throughout. Therefore, manipulation of the width or quality of a visual image can directly affect the population activity of visual neurons. The auditory periphery, however, is tonotopically organized—hairs cells in the cochlea are organized according to the sound frequencies they encode and do not directly encode sound-source location. Therefore, the auditory system must estimate the location of sound sources on the basis of interaural differences of arrival time and intensity (ITDs and ILDs) as well as on the spectral characteristics imposed by the HRTFs—(see Knudsen and Brainard 1995; Middlebrooks et al. 2002, and also see Sect. 2 above). The auditory brainstem extracts these localization cues in computations that involve multiple neural structures. The resulting localization cues do not always unambiguously correspond to a single physical sound-source location but rather to a locus of possible locations—the “cone of confusion”—see Sect. 2. The computational nature of auditory space means that “blurring” an auditory image is not as straight forward as it is for a visual stimulus. Perhaps as a result, multisensory research has only seldom manipulated the reliability of auditory localization cues.

However, studies have shown that poorly localized auditory stimuli tend to facilitate visual dominance. Thurlow and Jack (1973) found that the relatively poor acuity of auditory localization in the vertical plane resulted in a stronger ventriloquist effect than in the horizontal plane. Similarly, Spence and Driver (2000) found that ventriloquism was more likely for sound stimuli that are difficult to localize (e.g., a 2-kHz tone from multiple speakers) than for sound stimuli that are readily localized (such as white noise from one loudspeaker). The reliability factor explored in these investigations is related to the quality or width of an internal, neural estimate of the auditory event, not the quality or width of the physical stimulus, as in vision. Therefore the nature of the poor localizability is not straightforward to predict. Erroneous auditory localization could be caused by reduced resolution of a wide excitation pattern across many spatial channels or interaural-cue computation in a single spatial channel, or both. To our knowledge, the neural mechanisms for the saliency of auditory spatial perception remain largely untested.

Montagne and Zhou (2016) investigated whether manipulations of the congruence between ITD and ILD affects the reliability of auditory responses and the magnitude of visual bias. Broadband noise bursts (15-ms duration) were presented from two hidden loudspeakers at $\pm 45^\circ$ about the midline, with or without a simultaneously presented light-emitting diode (LED) flash from -45° , 0° , or $+45^\circ$.

Two auditory conditions were contrasted, (1), timing-based stereophony with incongruent ITDs and ILDs and, (2), level-based stereophony with congruent ITDs and ILDs. Figure 3 shows the relationship between the standard deviation (SD) of auditory-alone responses and the change in auditory localization when the light stimulus was present, that is, the visual bias, ΔAV . Listeners localized sound sources with greater variability and stronger visual bias for the timing stimuli than for the level stimuli. Also, the magnitude of visual bias for the timing signals correlated strongly with the variance (noise) of listeners’ auditory estimate, suggesting an intrinsic link between binaural ambiguity and localization uncertainty. In turn, the putative

Fig. 3 Relationship of response variability and visual capture for individual subjects. Symbols indicate (average) responses of individual subjects for timing-based stereophony (**squares**), level-based stereophony (**circles**) and single-speaker controls (**asterisks**). Straight lines show linear fits for each condition. From Montagne and Zhou (2016)



uncertainty of auditory localization modulated the strength of visual bias on sound localization.

3.3 Asymmetry of Perceptual Space

When the head and body are stationary, the visual and auditory systems do not encode the same spatial range. Auditory space is broad and extends to both front and rear space, whereas human vision is restricted to the frontal region, with visual acuity declining towards peripheral locations away from the fovea (Curcio et al. 1990). The resulting asymmetry between visual and auditory space is an important factor to consider in addition to the differences between the peripheral mechanisms in vision (spatiotopic encoding) and audition (computational space based on tonotopic encoding). Despite these differences, our knowledge of cross-modal spatial bias is mostly limited to audio-visual (AV) interactions in the frontal hemifield. As mentioned earlier, the symmetry of interaural cues along sagittal planes normal to the interaural axis often leads to front-back reversals. Indeed, the question of whether frontal visual cues can interact with the auditory events that are perceived in the rear, be they real or illusory, remains an interesting and ecologically important research topic.

Montagne and Zhou (2018) investigated the influence of frontal LED flashes on the perceived front-back, left-right location of a phantom sound source generated using timing-based stereophony. Figure 4 shows that there was a considerable amount of front-back confused responses to a center-position phantom source presented either from front or back. The colored lines show that frontal visual cues increased the percent of frontal responses. Left-right response shifts can be seen to follow the direction of the light. Interestingly, the lateral visual bias is only observed for the perceived frontal sound sources at 0°. Very little lateral bias was found in the perceived sound sources at 180°. The study also revealed that increasing the stimulus duration reduced both the rate of front-back reversals and the visual bias but not

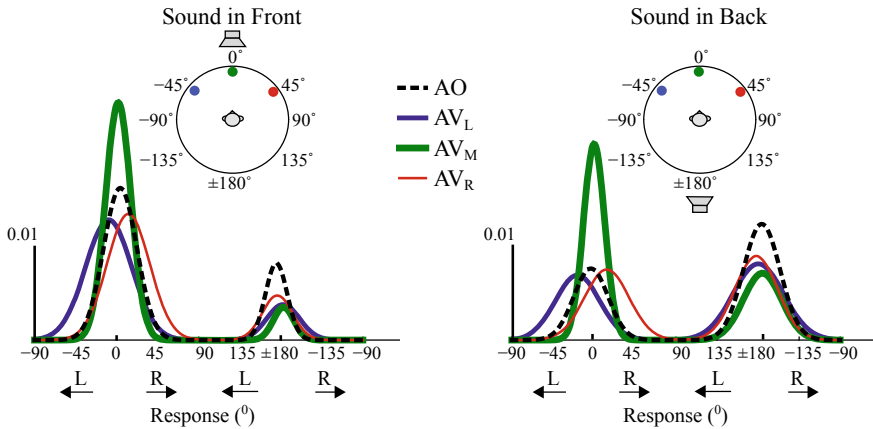


Fig. 4 Two-peak Gaussian functions for *AO*, audio only, and *AV*, audiovisual, responses for 15-ms duration Gaussian noise stimuli. The left and right figures show the results obtained using the two frontal or rear speakers, respectively. Each curve was obtained by fitting the data from all trials and all subjects using the Gaussian-mixture model. The delay between the two loudspeakers was 0ms for both conditions shown. The speaker sign marks the expected position of the perceived “phantom sound source.” The *AO* results (**black dashed line**) show that the responses were clustered on the midline at 0° and ±180°. The colored lines show changes in the left-right and front-back responses after adding visual stimulation. From Montagne and Zhou (2018)

localization errors associated with left-right judgment. These findings show that visual information separately interacts with left-right and front-back dimensions of a perceived sound source, while stimulus duration mainly modulates front-back errors in multisensory spatial processing.

The interactions between frontal vision and rear audition do not easily fit with existing Bayesian statistical models (e.g., Ernst and Banks 2002; Battaglia et al. 2003; Alais and Burr 2004) because these models are primarily based on the results of cross-modal perception of a seen target. In other words, the stimulus, S , in the prior distribution, $p(S)$, has an implicit frontal origin. Furthermore, the modality-specific, sensory representation, likelihood estimate, $p(R_A/S)$ or $p(R_V/S)$, consists of a one-to-one mapping of S in the form of a unimodal (single-peaked) likelihood function. As shown in Fig. 4, this estimate is not adequate after considering the rear sound field, where the front-back confused responses result in a bimodal likelihood function, $p(R_A/S)$. These factors complicate the variance estimate and subsequently the construction of the posterior probability using combined auditory and visual estimates as shown in Eq. (3).

Montagne and Zhou (2018) suggested an alternative mode of AV interaction for when the stimulus space extends outside the field of vision. They proposed that visual processing might affect the left-right and front-back auditory judgment independently in two different stages, (1), an initial coarse and broad auditory detection to decide the relative front vs. back direction of an event and, (2), if the perceived target location is in front, visual analysis to refine the estimate using integrated auditory and visual information. According to the causal-inference theory, the brain should limit the

extent of integration between sensory events perceived to rise from different sources (Kording et al. 2007). Montagne and Zhou (2018) argued that the causality test likely occurs during the initial auditory detection stage, which includes front-back discrimination.

4 Sound-Source Localization with Moving Listeners and/or Moving Sound Sources

Section 3 showed evidence for the integration of head-centric auditory spatial estimates with world-centric visual estimates under conditions where listeners and sound sources were stationary. This section considers evidence from scenarios where listeners and/or sound sources move, especially with sound stimuli that offer no spectral cues to specify where a target sound source is on a given cone of confusion. Wallach (1939, 1940) has been continuously cited in the literature with regard to the role head motion plays in avoiding front-back reversals. However, Wallach's foundational insight that multisensory, multi-systems information about head position must be integrated with interaural-difference cues in order to localize sound sources to their position in the surrounding environment, has received little attention until very recently.

This section, therefore, begins by reviewing some of Wallach's experiments and the logic that inspired them. To begin with, the simplest case for Wallach's hypothesis, that listeners could resolve spatial ambiguities in the azimuth plane by using head movements to compare the change in head-related auditory cues to the change in head position, is examined. The section then considers how Wallach extended this insight to propose a possible mechanism for estimating the elevation of sound sources without using spectral cues. Current knowledge about the head-position cues that might be integrated with the interaural cues in determining world-centric sound-source location is then reviewed. Finally, there is a brief review of some current investigations of the integration of interaural and head-motion cues.

4.1 *The Wallach Azimuth Illusion*

Wallach (1938, 1939, 1940) noted that interaural difference cues alone (especially ITDs) specify not just a single location, but an entire locus of positions, a "cone of confusion," all with the same angular relation to the head—see Sect. 2.2 for further details. As Wallach (1939) showed, head movements can be used to determine the front/back location of a stationary sound source—see Fig. 5. Wallach hypothesized that the relation between the change in interaural difference cues, relative to a given change in head position, would allow listeners to reduce the cone of confusion to a single point, thereby avoiding front-back reversals. An essential component of this hypothesis is that the listener makes some assumption about the movement, or

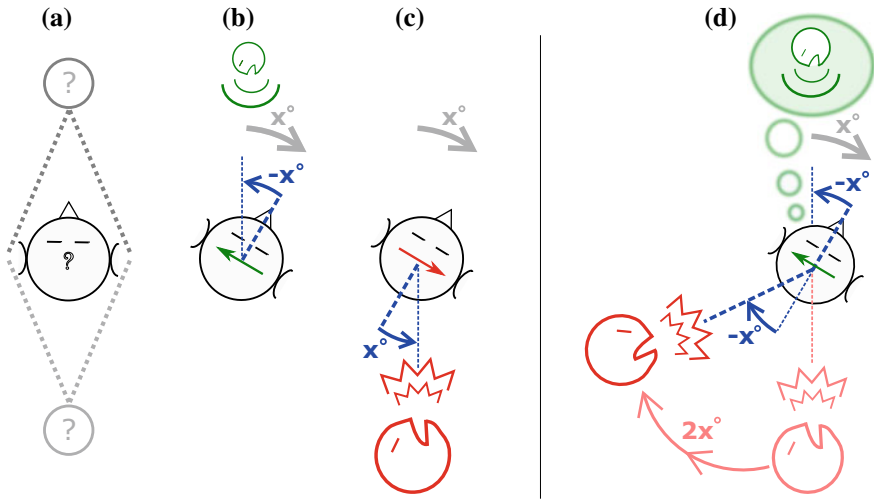


Fig. 5 Left (a, b, c) The basic idea of how head movements can be used to disambiguate front-back confusions. For world-centric changes of sound-source position (**red arrows**) and head position (**gray arrows**), clockwise rotation is notated as positive. For head-centric interaural differences, changes (**blue arrows**) that favor the right ear are notated as positive. Column (a) shows that low-frequency interaural differences, especially ITDs, are the same regardless of whether the sound source is in front or in back of the listener. (b) For a sound source in front of the listener on the azimuth plane at 0° elevation, a head turn of x° (**grey arrow**) results in a change in interaural differences equivalent to a $-x^\circ$ (**blue arrow**) change in sound-source position. (c) The same head turn results in a change in interaural cues equivalent to x° (**blue arrow**) for a sound source behind the listener. (d) A visual explanation of Wallach’s Azimuth Illusion. By rotating a sound source at twice the rate of the listener’s head turn of $2x^\circ$ (**red**), the same change in interaural-difference cues that would occur for a stationary sound source in the opposite front/back hemifield (front, in this case: $-x^\circ$) is produced. Provided there are no spectral cues to disambiguate front from back, the listener hears a stationary sound source in front, at the location of the green sound source, even though the actual (**red**) sound source is moving behind the listener at twice the listener’s rate of rotation. Note that the difference in sign between world-centric and head-centric angular rotation highlights the disconnection between the two coordinate systems that must somehow be bridged

lack thereof, of the sound source during the head movement. Specifically, Wallach assumed that “of all the directions which realize the given sequence of lateral angles, that one is perceived which is covariant with the general content of the surrounding space.” That is, assuming the sound source is stationary, there will only be one point in space, at or above the height of the pinnae, that is common to all cones of confusion that exist along the trajectory of the listener’s head rotation—the *Selective Principle of Rest*.

To test this notion, Wallach (1939, 1940) created an experimental apparatus that was coupled to the listener’s head. The device had electrical switches that activated, as a function of the listener’s head movements, one of 20 equidistantly spaced loudspeakers on a 120° circular arc. Wallach calculated the rate at which the head-centric auditory spatial estimate, derived from interaural-difference cues, would change dur-

ing a head turn for a sound source in front of the listener. He then produced the same changes in sound-source location (relative to the listener's head) that would occur for a frontal sound source. However, he presented the sound from *behind* the listener, rotating at twice the rate of the listener's head rotation. Figure 5D offers a graphical explanation of this basic concept—see Sect. 4.2 and Yost et al. (2019) for more detailed, mathematical explanations. Given a stimulus conducive to front-back reversals, the listener hears a stationary sound source in the front-back hemifield opposite to the one from which the stimulus was initially presented, despite the fact that the sound source is actually rotating around the listener in the same direction but at twice the rate of the listener's rotation. This suggests that the listener determines the front/back location of the sound source using the concomitant changes in interaural-difference cues for a given head turn. Since the change in interaural cues is commensurate with the magnitude of the head turn, the listener assumes the sound source is static. This basic result was reported by Wallach (1940) for all five tested listeners. For a review of perceived auditory motion, see Carlile and Leung (2016).

There are at least two possibilities for how the Wallach Illusion, and dynamic world-centric localization in general, could occur. It is possible that the world-centric location of auditory objects is updated at relatively sparse intervals, and that localization is head-centric between these intervals. For example, localization could be world-centric before and after a head turn, but head-centric during the turn due to the increased complexity and reduced resolution of dynamic sound-source localization. Following this notion, the change in interaural cues would be compared with the change in head position. The result of this comparison would then be mapped to world-centric coordinates for a “spatial update.” Under such conditions, one might expect vestibular cues to provide useful information regarding the change in head position in between spatial updates. Another, perhaps more computationally intensive possibility, is that the auditory system continuously updates world-centric coordinates of a perceived sound source. In this case, changes in the world-centric estimate(s), which could be bimodal if the possibility for front-back reversals exists, or even a locus of possible source positions in the form of a cone of confusion, would be compared with the head position. The comparative trajectory of the sound source and head position estimates would then determine the singular estimate of the sound-source position in the local environment. Targeted experiments will be required to reveal which of the two hypothesized processes is more appropriate—(see also Brimijoin and Akeroyd 2014, reviewed below).

4.2 *The Wallach Vertical Illusion*

The direction-dependent filtering provided by the pinnae, head, and torso—the so-called head-related transfer function (HRTF)—may not be the only elevation/front-back cue. Wallach (1938, 1939, 1940) extended his Azimuth Illusion—see Sect. 4.1—to include the judgment of elevation, pointing out that the rate at which interaural cues change relative to head motion could be used, assuming a stationary

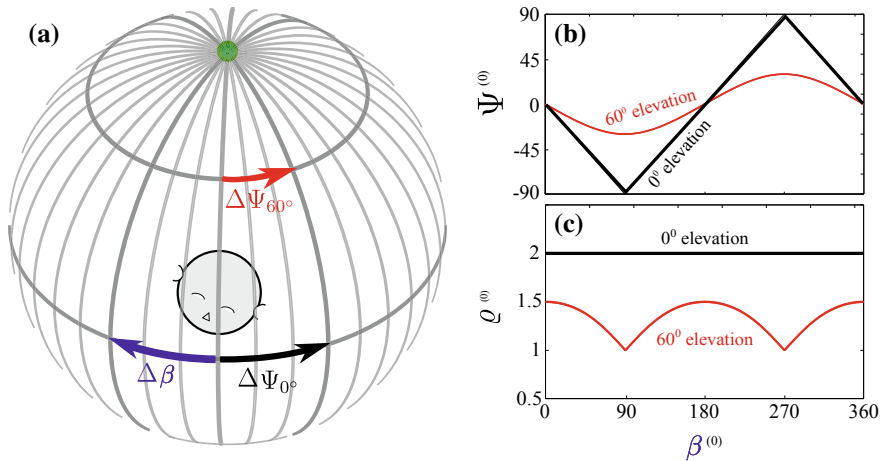


Fig. 6 a Visual description of Wallach’s basic concept. For a stationary stimulus on the azimuth plane, that is, with 0° elevation, a listener’s head rotation, $\Delta\beta$, results in the same but opposite change in angular displacement of the sound source relative to the head as would occur if the sound source had traveled the same angular distance in the opposite direction, $\Delta\Psi_{0^\circ}$. If the sound source is at an elevation ν that is off the azimuth plane, for example, Ψ_{60° , then the change in interaural-difference cues will be less for the same head movement, $\Delta\beta$. For a sound source above the head (green dot), there are no changes in interaural differences for a given head turn. **b** The lateral angle, Ψ , as a function of head position, β , where the frontal midline is 0°. **c** The ratio, ρ , of sound-source rotation relative to listener head rotation required to induce the Wallach Illusion, is shown as a function of the head position, β

sound source, to determine elevation of a sound source in the absence of spectral HRTF cues.

Figure 6a illustrates Wallach’s basic insight. The position of the head, β , is measured relative to the midline. Interaural differences are instead considered relative to the *interaural axis*, which can be imagined as a line passing through both ears. This is also the axis about which the *cones of confusion* are centered. Wallach called the angle between the sound source and the interaural axis the *lateral angle*, Ψ . This value corresponds to an interaural difference. Note that *both* azimuth and elevation contribute to the lateral angle, Ψ , in the following way:

$$\Psi = 90^\circ - (\cos^{-1}(\sin \beta \cos \nu)) . \tag{5}$$

That is, despite the common conception that interaural disparities are used only for encoding azimuth, a given interaural difference actually corresponds to a range of positions at many elevations—see Sect. 2.2 for further details. If a sound source is located anywhere on the sagittal plane corresponding to the midline, the interaural differences are approximately zero. Note that, in Wallach (1940), this condition would be notated as $\Psi = 90^\circ$. For this chapter, Ψ is reduced by 90°, so that the midline corresponds to $\Psi = 0^\circ$. Therefore, the “lateral angle” in this case is really

the displacement of the sound source from the median plane instead of the interaural axis.

Turning again to Fig. 6a, our listener makes a head turn of β . If the sound source lies on the 0° elevation azimuth plane, the corresponding change in Ψ (indicated by the **black arrow**) relative to β is $\frac{\Delta\Psi}{\Delta\beta} = 1$, the same as β . This is the maximal change in Ψ for a given head turn. At the other extreme, a sound source directly above the listener will elicit the smallest possible change, $\Delta\Psi = 0$.

Wallach realized that for a sound source at some intermediary elevation, say $\nu = 60^\circ$, the change in Ψ will also be intermediary, as indicated by the red arrow in Fig. 6a. Figure 6b shows how Ψ , the angular relation of the sound source to the median plane, changes as a function of head rotation, β . Note that, at 0° elevation, there is a unity gain between β and Ψ , whereas for 60° sound-source elevation a change in β results in far less of a change in Ψ . The maximum value, $|\Psi|$ can take at any elevation is the complement of ν , for example, 30° for a sound source at 60° elevation—see Mills 1972 for a similar derivation.

Using this information, the ratio of sound-source rotation to head rotation which is necessary to induce the Wallach Illusion, ϱ , can be calculated for a sound source, presented from the 0° -azimuth plane, as

$$\varrho = \frac{\Delta\Psi}{\Delta\beta} + 1. \quad (6)$$

For a signal without sufficient high-frequency information to allow a listener to exploit pinna-based cues, a purely rotational head movement will not allow the listener to determine if a sound source is above or below the 0° azimuth plane. If even a small head tilt is included in the head movement, however, this ambiguity could also be avoided.

To test this, Wallach could have asked rotating listeners to judge the elevation of stationary sound sources. Instead, Wallach (1940) employed the same argument that leads to the Wallach Azimuth Illusion to show how azimuthal head rotation, coupled to azimuthally-rotating sound sources, could lead to the illusory perception of a stationary sound source at an elevation specified by the speed of sound-source rotation relative to the listener's rotation, ϱ . In his main experiment, Wallach simulated a sound source at an elevation of 60° , above the horizontal plane. He did so by rotating a listener passively sitting in a chair (either blindfolded or not) with the sound source rotated at 1.5 times the rate of head rotation from behind the listener. This rate of rotation, an approximation to (6), was expected to induce a perceived elevation angle of $\nu = 60^\circ$, given that the listener only rotated within a relatively narrow angular range. Fifteen listeners indicated that the musical sounds were perceived above them in elevation, more so when their eyes were open than when they were closed. However in many cases, the listeners' judgments of elevation underestimated the predicted elevation of 60° .

Since Wallach's (1940) calculations only indicate a change in elevation relative to the horizontal plane and not whether the vertical angle is positive (above the pinnae) or negative (below the pinnae), a response below the horizontal plane would be consistent with his calculations. In this regard, Wallach (1940) made two some-

what inconsistent assumptions. First, he assumed that listeners' experience naturally biased them to perceive sounds above them rather than below them. Second, Wallach argued that perceived sound-source locations below the listener may have influenced listeners to underestimate elevation. It is worth noting that, at elevations other than directly above/below the listener or on the 0° -elevation azimuthal plane, the rate of change in interaural cues for a given head rotation is not constant but rather essentially a rectified sinusoidal function—see Fig. 6c. Thus, another possibility is that the linear estimation of the rate of sound-source rotation was too coarse an approximation to elicit the full illusion.

4.3 *What Are the Cues for Head Position?*

Both the horizontal and vertical illusions reported by Wallach (1940) suggest that head motion is a crucial variable in sound-source localization. Wallach (1940) assumed that “three types of sensory data represent a displacement of the head, that is, proprioceptive stimulation from the muscles engaged in active motion, stimulation of the eyes, and stimulation of the vestibular apparatus.” In this section, some of the current knowledge regarding these and other possible head motion cues is reviewed.

Clearly, vision provides an important estimate of head position—except when our eyes are closed. Previous visual experience is nevertheless likely to be useful even with eyes shut (Zwiers et al. 2001)—see Sect. 3. Head and eyes often move independently, and nearly constant eye movements could make the formation of a stabilized image of the outside world impossible. To cope with this, the visual system employs an eye-centric reference system in addition to a head-centric reference system. To stabilize perception of visual objects, the vestibulo-ocular reflex (VOR) and the optokinetic reflex (OKR) work together to provide a means to correct the retinal output for retinal movement. There is some evidence that, in addition to a head-centric reference system, an eye-centric reference system that involves eye motion and sound-source localization may also play a role in sound-source localization—see Van Opstal (2016).

The vision literature shows that head-position signals can be used to “correct” spatial visual cues by use of efferent (efference) copies or corollary discharge signals—see Van Opstal (2016). The general idea is that when a neural signal is generated to control head position, a copy (*efferent copy or corollary discharge*) is also made. This copy is then integrated with the retinal spatial signal to yield a stable perception of the world. For instance, if there is a stationary light source and the head moves, the retinal output would change. The efferent copy/corollary discharge would indicate that it is the head that moved and not the light source. This efferent-copy signal could be used to effectively cancel the retinal change signal, yielding a veridical estimate of the location of the stationary visual source. In the visual literature, there are several well-established examples of such a “cancellation” based on both eye movements and head movements (Bridgeman and Stark 1991).

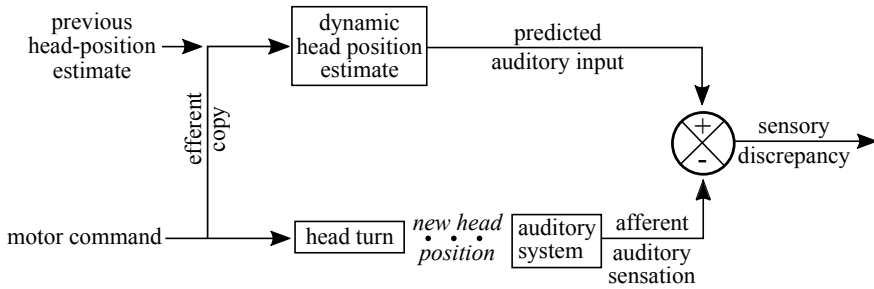


Fig. 7 A simplified schematic model for a possible role of an efferent-copy process in auditory localization

Figure 7 offers a simplified schematic of how efferent copy might work with auditory localization, based on an *efferent-copy* process in the visual system. Before a listener rotates the head, an internal estimate of the head position already exists. A motor command sends the signal to neck muscles and other involved systems to turn the head. Another “copy” of this signal is sent elsewhere in the brain so that a series of new, dynamic head position estimates can be made. Based on these estimates, the change in auditory spatial cues and/or auditory spatial estimates can be calculated. At approximately the same time, new afferent auditory activity offers a new spatial estimate of sound-source position that can be compared to the predicted auditory output, allowing the listener to determine whether the sound source has remained stationary or, if not, the position to which it has moved. While there is no direct physiological evidence for such efferent copy/corollary discharge processes in the mammalian auditory system, several authors (e.g., Wallach 1940; Brimijoin and Akeroyd 2012; Genzel et al. 2018; Freeman et al. 2017) have suggested such processes for sound-source localization.

The vestibular system also offers cues for determining a change in the head position. Vestibular cues result from the head’s angular acceleration, which triggers hair-cell responses in the semi-circular canals that in turn elicit neural impulses to inform an estimate of head-position (Lackner and DiZio 2005). Because the otoliths in the vestibular system act as accelerometers, there is no vestibular output when the head is kept still, nor is there any output when the head rotates at constant velocity. In most experiments and most everyday experience, both passive and active listener rotation include changes in velocity—self-rotation necessarily includes acceleration and deceleration. Yost et al. (2015) appears to be the only study in which sound-source localization judgments were partitioned according to whether listeners rotated in an accelerating, decelerating, or constant manner—compare see Sect. 4.4 for details.

The arguments presented in this chapter imply that having access to the head-position angle is important to establishing a head-position cue. It is worth noting that the vestibular system provides information that the head is rotating, the direction of rotation, and the relative velocity of rotation, but the vestibular system cannot by itself indicate the world-centric position of the head since it directly encodes only

the change of the position of the head. The absolute head angle could be computed if the time over which the rotation had occurred and the starting head location were known—this would, however, require memory and other sensory inputs. This idea, in terms of establishing head-position cues for world-centric sound-source localization, appears to be unexplored.

Several sound-source-localization studies either indirectly infer or directly implicate proprioception and/or neural motor control of head rotation as ways to gain information about the current head position. In most cases, ideas from the vision literature are used to infer how proprioceptive outputs could inform head-position cues. When listeners move, neural-motor control signals are required to initiate and control the movement. These signals could indicate the angle of the head. In addition, it is possible that when listeners are rotated by some external means and must keep their heads still, that resistance to the rotational motion would stimulate muscles (e.g., neck muscles) which would trigger neural signals as a means of indicating head rotation. However, it isn't clear how such resistance would inform the estimate of head-position angle, nor is there much physiological evidence for how such proprioceptive/neural control signals interact with physiological sound-source localization processes.

The other three possible processes that might provide head-position cues, namely, auditory cues from sound sources other than the “target” stimulus, somatosensory cues, and cognitive processes (spatial maps) have not been studied as far as the authors can tell. It seems logically possible that these cues could inform the spatial system about head position and thereby contribute to sound-source localization—they should thus be investigated.

4.4 Recent Studies of Sound-Source Localization as a Multisensory Process

While Wallach's research is seminal in establishing a multisensory approach to understanding sound-source localization, there are several aspects of his work that need to be considered in light of current relevant knowledge. First, Wallach (1940) presented music played by a Victrola record player. Due to the constraints of the technology of the time, this likely means that the sound stimuli were essentially low-pass filtered, removing any useful HRTF/pinna cues (note that noise from scratches and dust on the record would also be filtered in the same way). This resulted in listener performance that led Wallach to believe the “pinna factor” was likely subservient to the integration of changing interaural cues with changing head position. Later experiments, reviewed below, would show that HRTF cues can remediate front-back reversals so that the listener hears the rotating sound source circling around the azimuth plane, and the Wallach Illusion fails.

Second, Wallach manually rotated listeners in a swivel chair back and forth over an arc of approximately 60°, with the eyes closed and the head fixed in a head holder.

Wallach also ran experiments with a rotating visual screen that induced the sensation of listener motion in the direction opposite to the screen's rotation to show that the Wallach Illusion could also be induced without listener movement, provided the listener received the same visual stimulation as would accompany a head movement (c.f., McAnally and Martin 2008). Unfortunately, the relative weightings of the different sensory and systems inputs were not measured in Wallach's experiments.

Perrett and Noble (1997a, b) attempted to replicate Wallach's elevation experiments. However, they were only able to replicate Wallach's findings for low-frequency sounds, suggesting that when reliable HRTF cues are present in the stimulus they override elevation cues derived from listener motion. For low-frequency stimuli, the correspondence between the predicted elevation and the actually judged elevations was only approximately 2/3 of the target elevation. Given the limited acuity of dynamic sound-source localization together with "binaural sluggishness," this result is not altogether surprising. Indeed, auditory resolution of elevation along the midline is also considerably poorer compared to localization on the azimuth plane.

Thus, until there are additional data, the current literature suggests that elevation cues provided by head motion are subservient and considerably less useful than HRTF spectral information for judging elevation. However, this may not be the case for machine listening. Zhong et al. (2016) used the Wallach concept to show that machine-learning algorithms (e.g., Kalman filters) could learn to use simulated head motion to determine the location of up to three different simultaneously presented sound sources located in different azimuthal and vertical locations.

Early work relating to head movement for the avoidance of front-back reversals and judging elevation can also be found in the 1938 thesis of Alva Wilska—see Kohlrausch and Altosaar (2011) and de Boer and van Urk (1941), also referenced in Blauert (1997).

Macpherson (2011) was interested in the relative weighting of spectral cues versus dynamic interaural differences in resolving front-back reversals. He designed an analogous version of the Wallach-Azimuth-Illusion experiment in a virtual auditory space, whereby he presented stimuli with various center frequencies and bandwidths. Data from only one listener have been reported. They indicate that when the stimulus was a low-pass noise (0.5–1 kHz), so that spectral cues were not available, listeners perceived a static sound source, front-back reversed to where it had originally been presented—as in Wallach (1940). Macpherson (2011) also tested narrow-band, high-frequency-noise stimuli and found that the Wallach Illusion failed. Macpherson thus suggested that this result could indicate that ILDs may not provide a sufficient basis for the dynamic auditory processing required for the Wallach Illusion. It should be noted, however, that Macpherson (2011) presented stimuli from in front of the listener, so that listeners would have to confuse a frontally-presented stimulus for one presented from behind. However, it has been repeatedly demonstrated that listeners tend to localize narrow-band, high-frequency stimuli to the frontal hemifield, independent of the actual location of presentation—e.g., Blauert 1969, 1997; Morimoto and Aokata 1984; Middlebrooks et al. 1989; Middlebrooks 1992. It may therefore be the case that the so-called "directional bands" are implicated in this result.

Brimijoin and Akeroyd (2012, 2017) also investigated the Wallach Azimuth Illusion. In their experiments, normal-hearing and hearing-impaired listeners moved their heads back and forth between $\pm 15^\circ$ of the midline. A camera system recorded the head motion and the system's output controlled amplitude panning of the sound such that the location of the *phantom sound source* was at twice the angle of the listener's head angle, thereby generating the "2-1" rotation necessary for the Wallach Azimuth Illusion. A low-pass filtered speech signal was presented and the filter cutoff was raised from 500 Hz to 16 kHz between conditions in octave steps. As listeners started to rotate their heads, a moving speech sound was either presented from a loudspeaker directly in front, or from a loudspeaker directly behind the listener. Listener responses indicated that they perceived a stationary sound source in the hemifield opposite to where the rotating sound was first presented. However, listeners responses were less robust in terms of replicating the Wallach Azimuth Illusion as the speech sounds included more and more high-frequency information. The authors state that "signals with the most high-frequency energy were often associated with an unstable location percept that flickered from front to back as self-motion cues and spectral cues for location came into conflict," perhaps suggesting that the brief duration of the presentations did not allow for the listeners to fully experience rotation.

Pastore and Yost (2017) and Yost et al. (2019, 2020) conducted an experiment that was an approximate replication of Wallach's (1940) study, but with a different means of rotating the listener and sound sources. The rate of front-back reversals (FBRs) was measured for noise stimuli under static listener/sound source conditions. Listeners were then rotated via a computer-controlled chair at a constant velocity of $45^\circ/\text{s}$. The sound-source rotated at twice the rate of listener rotation by way of saltatory motion from loudspeaker to loudspeaker around a circular array consisting of 24 equally spaced (15° apart) loudspeakers. Five differently filtered 200-ms noise bursts were tested, namely, three that generated more than 35% FBRs (FBR likely) and two that generated fewer than 6% FBRs (FBR unlikely)—for further details, see Fig. 8.

After eight seconds of stimulus presentation, the listener indicated the direction of rotation (clockwise or counterclockwise) for stimuli perceived as rotating, or the loudspeaker (separated by 60° , the same as in the first experiment) that most closely corresponded with the perceived static sound-source location.

Figure 8 depicts the effects of the stimulus spectrum and whether the listeners' eyes were open or shut. The left two panels show results when seven listeners' eyes were open, giving them information about the head position. The right two panels are the results from six of the same seven listeners when the listeners' eyes were closed—in a dark room and wearing a blindfold. One might expect that, in the eyes-closed condition, listeners have little or no access to information about the position of their head, thereby restricting their localization to the angular relation of the sound source to the head. In this case one would expect listeners to perceive a rotating sound source with their eyes closed, regardless of the stimulus frequency—see Yost et al. (2015) for more details about the assumptions regarding head-centric versus world-centric sound-source localization when attempts are made to eliminate head-position cues.

For the filtered noises that were prone to FBRs (FBR likely), listeners perceived the sound as being stationary when the eyes were open (consistent with the Wallach

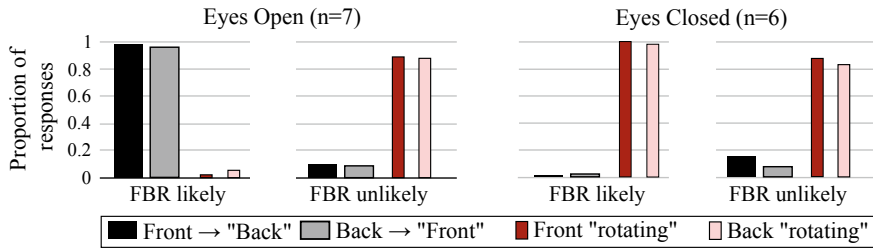


Fig. 8 Data from Pastore and Yost (2017). Results are pooled across noise stimuli that were likely to elicit front-back reversals (“FBR likely”)—250-Hz center frequency, 2-octave and 1/10-octave bandwidths—and unlikely to elicit front-back-reversals (“FBR unlikely”)—4-kHz center frequency, 2-octave and 1/10-octave bandwidths. At the beginning of either listener or sound-source rotation, the sound stimulus was presented from directly in front or from directly behind the listener. The listeners indicated whether the sound was perceived at a fixed location or whether it was rotating. For example, the proportion of those responses is shown by the black bars, for which stimuli that were presented either in front or from behind were indicated to be stationary

Azimuth Illusion), but rotating when the eyes were closed. For the filtered noises that are not prone to FBRs (FBR unlikely), listeners nearly always perceived the sound as rotating in both the eyes-open and eyes-closed cases, indicating that the Wallach Illusion fails for stimuli that are not prone to FBRs. In the eyes-open condition with the listener facing forward, the perception of a stationary noise source was nearly 100% of the time at the rear loudspeaker when the sound was presented from the front, and at the ‘frontal loudspeaker when the sound was presented from behind. When listeners’ eyes were closed and the sounds were not likely to produce FBRs, listeners always indicated that the sound rotated clockwise—as the actual sound did. When the eyes were closed and the noise was likely to elicit FBRs, listeners indicated that the sound was rotating in a clockwise direction most of the time, but occasionally counterclockwise rotation was also indicated. However, one listener’s responses in the eyes-closed condition when FBRs existed was not consistent with the other five listeners’ responses. Thus, listeners’ perception with their eyes closed, needs further investigation.

Brimijoin and Akeroyd (2014) studied the moving minimum-audible angle (MMAA), that is, the minimum-perceivable angle between two sound sources when the angular displacements of two sound sources change relative to a listener’s head. They reported that, when listeners rotated their heads and the sound sources were stationary, the MMAA was 1–2° smaller than when listeners kept their heads still and the sound sources rotated around the listener with the same angular velocity and displacement as the listeners’ previous head turns. Brimijoin and Akeroyd (2014) concluded that “spatial processing involves an ongoing and highly accurate comparison of spatial acoustic cues with self-motion cues.”

Brimijoin (2018) showed that the perceived motion of a moving sound source differs depending on its angular displacement. Sounds to the sides of listeners needed to be moved more than twice as far as sounds near midline for both sounds to

appear to have moved the same amount. How this relative compression/expansion of auditory space interacts with head position cues is unknown. One possibility is that the comparison of the two rates of motion is not very precise. Another possibility is that other inputs are employed, such as vision, to compensate for these distortions of auditory space, as considered in Sect. 3.

Yost et al. (2015) investigated several aspects of sound-source location when listeners were rotated in a chair and their eyes were either open or closed. They argued that listeners had little or no information about the position of their head when they rotated in the chair at constant velocity, and, with their eyes closed, there were no visual cues. Under these conditions, listeners perceived stationary sound sources as rotating. When sound sources and the listener rotated at the same rate, listeners perceived a stationary sound source—entirely consistent with localization based on a head-centric reference system. When these listeners' eyes were closed and the rotation was accelerating or decelerating, the results were somewhat mixed. Yost et al. (2015) point out that there were possible confounds in their procedures, making it difficult to unambiguously determine the role of vestibular acceleration/deceleration cues in judging head position. How these cues might thereby influence sound-source localization is therefore also unclear.

Genzel et al. (2016) investigated “spatial updating,” the process of mapping the head-centric auditory estimate of sound-source position to the listener's spatial map of the surrounding environment using successive estimates of head position. In three different experimental conditions, blindfolded listeners were either, (1), asked to move their head according to a trained rotational trajectory, (2), passively moved along the same trajectory, or (3), counter-rotated as a function of head rotation, such that a given head rotation resulted in no change in head position relative to the surrounding environment. In a two-alternative forced-choice experiment, listeners reported whether they heard a test sound to the right or left of a previously presented reference sound. Listeners were most accurate when passively rotated and least accurate when they moved their own heads. Genzel et al. (2016) modeled the integration of head-centric auditory spatial inputs and world-centric head position information as a linear addition, dividing head-motion cues into vestibular cues and proprioceptive/efference copy cues, with visual inputs zeroed out due to the listener being blindfolded. They determined that both proprioceptive/efference copy and vestibular cues play a role in determining head position, but that vestibular cues are weighted more heavily. While there are several untested assumptions underlying their interpretations, their data clearly indicate support for the notion that sound-source localization depends on the integration of head-motion and auditory spatial cues, and that vestibular function and proprioception/efferent copy are possibly used as indicators of head position.

Wightman and Kistler (1999) investigated whether head movements could be used to disambiguate front/back sound-source localization along cones of confusion. The authors tested this under four scenarios: (1), no head movement allowed, (2), the listeners moved their heads, (3), the listeners did not move their heads, but the sound source was moved by the experimenter, and (4), the listener did not move their heads, but they themselves moved the sound source via key presses on a computer keyboard.

Wightman and Kistler found that head movements reduced the front/back errors to almost zero (see Braasch et al. 2013, for related modeling). Listeners also reduced front/back reversals to a minimum when they themselves moved the sound source via keyboard with no visual or vestibular feedback. No such benefit was found when the experimenter moved the sound source. This important finding suggests that mapping head-related sound-source localization to the local environment involves cognitive processing that uses whatever information and spatial estimates that appear to be useful.

Motion parallax is a powerful cue used in vision to judge relative distance (Steinman and Garzia 2000). Genzel et al. (2018) demonstrated the possibility that motion parallax might play a similar role in judging the relative distance of sound sources. Their main experiment used a virtual panning process to present two sounds (a low-pitched and a high-pitched sound) at two different panned (virtual) distances. With no motion of the sounds or the listeners, stationary listeners could not determine whether one sound was further away from the other, since all known distance cues were eliminated. When listeners moved, they were much better at discriminating the differences in panned distances than when the sound source moved and the listener was stationary. In other words, listeners could infer that one sound was closer than the other one by exploiting the perceptual effect that the near-panned source appeared to move faster than the far-panned one while the head was moving. This is consistent with the visual analogy. There was a small decrement in performance when the listeners were moved on a platform rather than moving themselves. This suggests that proprioceptive cues for self-motion are involved when judging sound-source distance.

5 A Concept for a Model

This section offers a descriptive model of sound-source localization, based on Wallach's (1940) insight that auditory spatial information must be integrated with an estimate of the location of the listener's head relative to the surrounding environment to provide an estimate of the location of the sound source in that environment. Because the range of possible inputs is large, and their respective temporal-processing speeds and parameter spaces are potentially very different, the model offered here is not yet actually implemented but rather of a conceptual kind. In particular, it does not yet specify details of how the various inputs to the model are combined and compared.

Two crucial points should be mentioned at the outset. First, full development of the model requires further studies of the multi-system/multisensory interactions that are involved in auditory localization and in the generation of dynamic, multisensory spatial maps. Second, the model is not yet available as a flowchart, because this would be too complex. In fact, the overall process is not simply feed-forward, but rather includes feedback and other interactions between system elements and sensory input/output—compare Chap. 1, this volume.

Also, there is, as of yet, no controlled experiment available to test such a model even if it were precisely specified. As the literature reviewed in this chapter shows, only small parts of the overall process can be investigated at a time and subsequently modeled. As such, the individual model structures may differ in kind. For example, audio-visual interaction, considered in Sect. 3, may be adequately modeled within a Bayesian framework. Whether interactions between memory, attention, motor processes, etc., can also be modeled in this way is unclear. Also, the putative *spatial map* may be a dynamic system of various spatial maps with different references along many dimensions. In other words, the proposed model concept of auditory localization as a multisensory/multi-system process is primarily intended to be a tool for orientation in a yet largely unknown territory.

This section uses a notation where Θ'_{ab} denotes an estimate (indicated by the prime) of angular displacement in polar space, Θ , relative to the frame of reference, a , and in terms of the type of input, b . It is worth noting that one cannot be entirely sure what all the *frames of reference* might actually be. There appear to be *head-centric* and *world-centric* frames of reference as illustrated in Fig. 1, but there may be *body-centric* and other frames of reference as well. For example, a listener with closed eyes may be able to point to the location of a sound source while not being able to place that location into the context of other objects in the room, for instance, for reaching out to grab a buzzing mosquito in the dark. Even this example still requires some internal map of, at least, the body. Nevertheless, the basic argument of the model concept is that the auditory estimate of the location of a perceived sound source within the context of the local environment (world-centric localization, Θ'_{w_A}), is determined by the integration of an auditory estimate of the sound source's location relative to the head based on auditory spatial cues, Θ'_{A_h} , with a multisensory/multi-systems estimate of the location of the head relative to the local environment, Θ'_{w_h} .

This descriptive model does not specify how Θ'_{A_h} and Θ'_{w_h} would be combined, but rather suggests that sound-source localization requires the integration of information—including, but not limited to, perceptual cues from several (perhaps many) neural systems. This includes cognitive processes such as experience and memory—compare Buzsáki and Llinás (2017). The model assumes that any such integration also involves an assessment of the reliability of the cues employed for each estimate. Furthermore, the model assumes that each cue estimate and each estimate resulting from the integration of those estimates introduces error, (ξ_i). For example, Θ'_{h_A} is determined by weighted integration of the auditory spatial cues, and Θ'_{w_h} is determined by a weighted sum of the multi-system head-position cues mentioned above. The weight, w_i , of any particular auditory spatial or head-position estimate would be proportional to the external noise of the cue, due to variability in the stimulus along the relevant dimension, together with an internal noise term, ξ_i , that arises from the variability inherent to neural processes in general. Combining these estimates to arrive at Θ'_{w_A} introduces further error, again due to internal noise. The initial auditory estimate Θ'_{h_A} relates sound-source position to the head as follows.

$$\Theta'_{h_A} \propto [w_\psi \Psi', w_v \nu'], \quad (7)$$

where Ψ' is an estimate of the angle of the sound source relative to the interaural axis, the *lateral angle* in Wallach's terminology—see Sect. 4.2. Ψ' is therefore a component of Θ'_{h_A} that is based on interaural differences of time, Ψ'_{ITD} , and level, Ψ'_{ILD} . Note that Ψ' is not simply an estimate of azimuth—the same interaural differences exist along a range of locations on the *cones of confusion* as discussed throughout this chapter. ν' is a polar elevation estimate on a sagittal plane normal to the interaural axis of Θ'_{h_A} , based on spectral HRTF cues. Therefore, *both* the component estimates, Ψ' and ν' , are required for a single-valued estimate of the sound-source location, Θ'_{h_A} . Furthermore, it is unclear whether elevation, ν' , can be estimated without an initial interaural-difference estimate, Ψ' , to specify which sagittal plane will be the basis for the elevation estimate,

$$\Theta'_{w_h} \propto [w_V V, w_\Gamma \Gamma, w_B B, w_A A, w_C C], \quad (8)$$

where $V \dots$ vision, $\Gamma \dots$ vestibular cues, $B \dots$ body awareness (e.g., proprioceptive, somatosensory, kinesthetic, neuro-motor control), $A \dots$ auditory, and $C \dots$ cognitive processes, which include expectation, memory, and attention. To determine an estimate of the sound-source position in the surrounding environment, the head-centric estimate, Θ'_{h_A} , must be combined with the estimate of head position, Θ'_{w_h} . This process could be analogous to simply adding the two estimates, or perhaps one is mapped onto the other—the actual mechanism is not understood at this time and, consequently, not specified in the following expression.

$$\Theta'_{w_A} \propto [\Theta'_{h_A}, \Theta'_{h_w}, C, \chi], \quad (9)$$

where χ denotes interactions between various estimates of Θ_w , such as auditory, Θ'_{w_A} , and visual, Θ'_{w_V} .

Several points are worth noting. First, although this model concept is expressed as a series of mathematical expressions, this form has only been chosen for convenience. The inputs and interactions between them for each of the spatial estimates are still largely unspecified. For example, the model could further include head-position cues that future research may suggest. The model is not expressed as addition of individual estimates since they may interact in non-linear ways. The model concept is meant to, hopefully, provide a structure to motivate experiments, the results of which could alter this putative model considerably. Second, the relative weighting of different sensory/systems input can be such that one or several are completely disregarded in a given estimate. For example, when listeners' eyes are closed, their visual input is probably not considered in any internal head-position estimate. Third, it is worth noting that Θ'_{A_w} is only one spatial estimate of a perceived sound source in the context of the surrounding environment among several other estimates from other sensory modalities, as well as from cognition. For example, if an estimate, Θ'_{V_w} , of the location of a visual object is perceptually grouped with the sound object associated with Θ'_{A_w} , these estimates will likely interact, either reinforcing each other or leading to cross-modal capture. Memory or expectation could play a similar role in this regard. This possibility is denoted by χ .

In summary, it is hoped that the contribution of the model concept is to point out that the roles in auditory localization as being played by several components of the assumed input, such as proprioceptive, somatosensory, kinesthetic, neuro-motor control, cognitive processing, and spatial auditory input used to determine head position, are still not clearly understood. Thus, the model primarily points to what remains unknown rather than at what is known.

Acknowledgements The work reported here is supported by the National Science Foundation (No. NSF BCS-1539376), the National Institute for Deafness and Communication Disorders (Nos. R0101DC015214 and F32DC017676), and Facebook Reality Labs. The authors are indebted to two anonymous reviewers for constructive comments and suggestions.

References

- Alais, D., and D. Burr. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* 14: 257–62. <https://doi.org/10.1016/j.cub.2004.01.029>.
- Anderson, P.W., P. Zahorik, J.A. Schirillo, and W. Forest. 2014. Auditory/visual distance estimation: accuracy and variability. *Frontiers in Psychology* 5 (October): 1–11. <https://doi.org/10.3389/fpsyg.2014.01097>.
- Battaglia, P.W., R.A. Jacobs, and R.N. Aslin. 2003. Bayesian integration of visual and auditory signals for spatial localization. *The Journal of the Optical Society of America* 20 (7): 1391–1397.
- Baumgartner, R., P. Majdak, and B. Laback. 2013. Assessment of sagittal-plane localization performance. In *The Technology of Binaural Listening*, ed. J. Blauert, 93–119. Berlin-Heidelberg-New York: Springer and ASA Press.
- Berkeley, G. 1709. An Essay towards a New Theory of Vision. <https://www.maths.tcd.ie/~dwilkins/Berkeley/Vision/1709A/Vision.pdf> (last accessed Dec. 20, 2020).
- Bernstein, L.R., and C. Trahiotis. 2011. Lateralization produced by envelope-based interaural temporal disparities of high-frequency, raised-sine stimuli: empirical data and modeling. *The Journal of the Acoustical Society of America* 129 (3): 1501–8. <https://doi.org/10.1121/1.3552875>.
- Bertelson, P., and M. Radeau. 1981. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception and Psychophysics* 29 (6): 578–584.
- Bertelson, P., J. Vroomen, B. de Gelder, and J. Driver. 2000. The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception and Psychophysics* 62 (2): 321–332.
- Best, V., D.S. Brungart, S. Carlile, N. Jin, E.A. Macpherson, R.L. Martin, K.I. McAnally, A.T. Sabin, and B.D. Simpson. 2011. A meta-analysis of localization errors made in free field. In *Principles and Applications of Spatial Hearing*, vol. 1, ed. Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and H. Kato, 14–23. Singapore: World Scientific Publishing. <https://doi.org/10.1142/7674>.
- Blauert, J. 1969. Sound localization in the median plane. *Acustica* 22, 205–213.
- Blauert, J. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*, 222–237. Cambridge: MIT Press.
- Boring, E.G. 1926. Auditory theory with special reference to intensity, volume, and localization. *The American Journal of Psychology* 37 (2): 157–188.
- Boring, E.G. 1942. *Sensation and Pareception in the History of Experimental Psychology*. New York: Appleton-Century-Crofts.
- Braasch, J., S. Clapp, A. Parks, M.T. Pastore, and N. Xiang. 2013. A binaural model that analyses aural spaces and stereophonic reproduction systems by utilizing head movements. In *The Technology of Binaural Listening*, vol. 8, ed. J. Blauert, 201–224. Springer and ASA Press.

- Bridgeman, B., and L. Stark. 1991. Ocular proprioception and efference copy in registering visual direction. *Vision Research* 31 (11): 1903–1913. [https://doi.org/10.1016/0042-6989\(91\)90185-8](https://doi.org/10.1016/0042-6989(91)90185-8).
- Brimijoin, W.O. 2018. Angle-dependent distortions in the perceptual topology of acoustic space. *Trends in Hearing* 22: 1–11. <https://doi.org/10.1177/2331216518775568>.
- Brimijoin, W.O., and M.A. Akeroyd. 2012. The role of head movements and signal spectrum in an auditory front/back illusion. *i-Perception* 3 (3): 179–181. <https://doi.org/10.1068/i7173sas>.
- Brimijoi, W.O., and M.A. Akeroyd. 2014. The moving minimum audible angle is smaller during self motion than during source motion. *Frontiers in Neuroscience* 8: 1–8. <https://doi.org/10.3389/fnins.2014.00273>.
- Brimijoin, W.O., and M.A. Akeroyd. 2017. The effects of hearing impairment, age, and hearing aids on the use of self motion for determining front/back location. *Journal of the American Academy of Audiology* 27 (7): 588–600. <https://doi.org/10.3766/jaaa.15101>.
- Bronkhorst, A.W., and T. Houtgast. 1999. Auditory distance perception in different rooms. *Nature* 397: 517–520.
- Brungart, D.S., N.I. Durlach, and W.M. Rabinowitz. 1999. Auditory localization of nearby sources. II. Localization of a broadband source. *Journal of the Acoustical Society of America* 106 (4): 1956–1968. <https://doi.org/10.1121/1.427943>.
- Buzsáki, G., and R. Llinás. 2017. Space and time in the brain. *Science* 358 (October): 482–485.
- Carlile, S., and T. Blackman. 2014. Relearning auditory spectral cues for locations inside and outside the visual field. *Journal of the Association for Research in Otolaryngology* 15 (2): 249–263. <https://doi.org/10.1007/s10162-013-0429-5>.
- Carlile, S., and J. Leung. 2016. The perception of auditory motion. *Trends in Hearing* 20: 1–19. <https://doi.org/10.1177/2331216516644254>.
- Choe, C.S., R.B. Welch, R.M. Gilford, and J.F. Juola. 1975. The “ventriloquist effect”: Visual dominance or response bias? *Perception and Psychophysics* 18 (1): 55–60.
- Curcio, C.A., K.R. Sloan, R.E. Kalina, and A.E. Hendrickson. 1990. Human photoreceptor topography. *Journal of Comparative Neurology* 292 (4): 497–523. <https://doi.org/10.1002/cne.902920402>.
- de Boer, K., and A.T. van Urk. 1941. Some particulars of directional hearing. *Philips Technical Review* 6: 359–364.
- Dorman, M.F., L.H. Loisel, S.J. Cook, W.A. Yost, and R.H. Gifford. 2016. Sound source localization by normal hearing listeners, hearing-impaired listeners and cochlear implant listeners. *Audiology and Neurotology* 21: 127–131.
- Ernst, M.O., and M.S. Banks. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415 (6870): 429–433.
- Ernst, M.O., and H.H. Bühlhoff. 2004. Merging the senses into a robust percept. *Trends in Cognitive Sciences* 8 (4): 162–169.
- Freeman, T.C.A., J.F. Culling, M.A. Akeroyd, and W.O. Brimijoin. 2017. Auditory compensation for head rotation is incomplete. *Journal of Experimental Psychology* 43 (2): 371–380. <https://doi.org/10.1037/xhp0000321>.
- Genzel, D., U. Firzlafl, L. Wiegrebe, and P.R. MacNeilage. 2016. Dependence of auditory spatial updating on vestibular, proprioceptive, and efference copy signals. *Journal of Neurophysiology* 116 (2): 765–775. <https://doi.org/10.1152/jn.00052.2016>.
- Genzel, D., M. Schutte, W.O. Brimijoin, and P.R. MacNeilage. 2018. Psychophysical evidence for auditory motion parallax. *Proceedings of the National Academy of Sciences* 115 (6): 4264–4269. <https://doi.org/10.1073/pnas.1712058115>.
- Good, M.D., and R.H. Gilkey. 1996. Sound localization in noise: the effect of signal-to-noise ratio. *The Journal of the Acoustical Society of America* 99 (2): 1108–17. <https://doi.org/10.1121/1.415233>.
- Goupell, M.J., and O.A. Stakhovskaya. 2018. Across-channel interaural-level-difference processing demonstrates frequency dependence. *The Journal of the Acoustical Society of America* 143 (2): 645–658. <https://doi.org/10.1121/1.5021552>.

- Hairston, W.D., M.T. Wallace, J.W. Vaughan, B.E. Stein, J.L. Norris, and J.A. Schirillo. 2003. Visual localization ability influences cross-modal bias. *Journal of Cognitive Neuroscience* 15 (1): 20–29.
- Hartmann, W.M., and B. Rakerd. 1989. On the minimum audible angle—a decision theory approach. *The Journal of the Acoustical Society of America* 85 (5): 2031–2041.
- Hartmann, W.M., B. Rakerd, Z.D. Crawford, and P.X. Zhang. 2016. Transaural experiments and a revised duplex theory for the localization of low-frequency tones. *The Journal of the Acoustical Society of America* 139 (2): 968. <https://doi.org/10.1121/1.4941915>.
- Hofman, P.M., J.G.A. van Riswick, and A.J. van Opstal. 1998. Relearning sound localization with new ears. *Nature Neuroscience* 1 (5): 417–421.
- Howard, I.P., and W.B. Templeton. 1996. *Human Spatial Orientation*, 359–362. New York: Wiley.
- Humanski, R.A., and R.A. Butler. 1988. The contribution of the near and far ear toward localization of sound in the sagittal plane. *The Journal of the Acoustical Society of America* 83 (6): 2300–2310. <https://doi.org/10.1121/1.396361>.
- Jack, C.E., and W.R. Thurlow. 1973. Effects of degree of visual association and angle of displacement on the “ventriloquism” effect. *Perceptual and Motor Skills* 37 (3): 967–979.
- Jackson, C.V. 1953. Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology* 5: 52–65.
- Jin, C., A. Corderoy, S. Carlile, and A. van Schaik. 2004. Contrasting monaural and interaural spectral cues for human sound localization. *The Journal of the Acoustical Society of America* 115 (6): 3124–3141. <https://doi.org/10.1121/1.1736649>.
- Jones, B., and B. Kabanoff. 1975. Eye movements in auditory space perception. *Perception and Psychophysics* 17 (3): 241–245. <https://doi.org/10.3758/BF03203206>.
- King, A.J. 2009. Visual influences on auditory spatial learning. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1515): 331–339. <https://doi.org/10.1098/rstb.2008.0230>.
- Knudsen, E.I., and M.S. Brainard. 1995. Creating a unified representation of visual and auditory space in the brain. *Annual Review of Neuroscience* 18: 19–43. <https://doi.org/10.1146/annurev.neuro.18.1.19>.
- Kohlrausch, A., and T. Altsaar. 2011. Early research on spatial hearing by Alvar Wilska (1911–1987). *Forum Acusticum*, 1103–1108. Aalborg: European Acoustics Association.
- Kolarik, A.J., B.C. Moore, P. Zahorik, S. Cirstea, and S. Pardhan. 2016. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, and Psychophysics* 78 (2): 373–395. <https://doi.org/10.3758/s13414-015-1015-1>.
- Kording, K.P., U. Beierholm, W.J. Ma, S. Quartz, J.B. Tenenbaum, and L. Shams. 2007. Causal inference in multisensory perception. *PLoS One* 2 (9): e943. <https://doi.org/10.1371/journal.pone.0000943>.
- Kuhn, G.F. 1977. Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America* 62 (1): 157–167. <https://doi.org/10.1121/1.381498>.
- Kuhn, G.F. 1987. Physical acoustics and measurements pertaining to directional hearing. In *Directional Hearing*, eds. W.A. Yost and G. Gourevitch, Chap 1, 3–25. Springer Nature. https://doi.org/10.1007/978-1-4612-4738-8_1.
- Lackner, J.R., and P. DiZio. 2005. Vestibular, proprioceptive, and haptic contributions to spatial orientation. *Annual Review of Psychology* 56 (1): 115–147. <https://doi.org/10.1146/annurev.psych.55.090902.142023>.
- Langendijk, E.H.A., and A.W. Bronkhorst. 2002. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America* 112 (4): 1583. <https://doi.org/10.1121/1.1501901>.
- Letowski, T., and S. Letowski. 2011. Localization error: accuracy and precision in auditory localization. In *Advances in Sound Localization, Chap. 4*, ed. P. Strumillo, 55–78. London: Intech Open. <https://doi.org/10.5772/597>.
- Macaulay, E.J., W.M. Hartmann, and B. Rakerd. 2010. The acoustical bright spot and mislocalization of tones by human listeners. *Journal of the Acoustical Society of America* 127 (3): 1440–1449. <https://doi.org/10.1121/1.3294654>.

- Macaulay, E.J., B. Rakerd, T.J. Andrews, and W.M. Hartmann. 2017. On the localization of high-frequency, sinusoidally amplitude-modulated tones in free field. *Journal of the Acoustical Society of America* 141 (2): 847–863. <https://doi.org/10.1121/1.4976047>.
- Macpherson, E.A. 2011. Head motion, spectral cues, and Wallach's 'principle of least displacement' in sound localization. In *Principles and Applications of Spatial Hearing, Chap. 9*, ed. Y. Suzuki, D. Brungart, and H. Kato, 103–120. Singapore: World Scientific.
- Macpherson, E.A., and J.C. Middlebrooks. 2002. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America* 111 (5): 2219. <https://doi.org/10.1121/1.1471898>.
- Makous, J.C., and J.C. Middlebrooks. 1990. Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America* 87 (5): 2188–2200. <https://doi.org/10.1121/1.399186>.
- Martin, R.L., M. Paterson, and K.I. McAnally. 2004. Utility of monaural spectral cues is enhanced in the presence of cues to sound-source lateral angle. *Journal of the Association for Research in Otolaryngology* 5 (1): 80–89. <https://doi.org/10.1007/s10162-003-3003-8>.
- McAnally, K.I., and R.L. Martin. 2008. Sound localisation during illusory self-rotation. *Experimental Brain Research* 185 (2): 337–40. <https://doi.org/10.1007/s00221-007-1157-z>.
- Mendonça, C. 2020. Psychophysical models of sound localisation with audiovisual interactions. In *The Technology of Binaural Understanding*. Springer, ed. J. Blauert, and J. Braasch, 289–314. Cham, Switzerland: Springer and ASA Press.
- Middlebrooks, J.C. 1992. Narrow-band sound localization related to external ear acoustics. *The Journal of the Acoustical Society of America* 92 (5): 2607–24.
- Middlebrooks, J.C., J.C. Makous, and D.M. Green. 1989. Directional sensitivity of sound—pressure levels in the human ear canal. *The Journal of the Acoustical Society of America* 86 (1): 89–107. <https://doi.org/10.1121/1.398224>.
- Middlebrooks, J.C., L. Xu, S. Furukawa, and E.A. Macpherson. 2002. Cortical neurons that localize sounds. *Neuroscientist* 8 (1): 73–83.
- Mills, A.W. 1960. Lateralization of high-frequency tones. *The Journal of the Acoustical Society of America* 32 (1): 132–134.
- Mills, A.W. 1972. Auditory localization. In *Foundations of Modern Auditory Theory*, ed. J.V. Tobias, 303–348. New York: Academic Press.
- Montagne, C., and Y. Zhou. 2016. Visual capture of a stereo sound : Interactions between cue reliability, sound localization variability, and cross-modal bias. *The Journal of the Acoustical Society of America* 140 (July): 471–485. <https://doi.org/10.1121/1.4955314>.
- Montagne, C., and Y. Zhou. 2018. Audiovisual interactions in front and rear space. *Frontiers in Psychology* 9 (MAY): 1–15. <https://doi.org/10.3389/fpsyg.2018.00713>.
- Morimoto, M. 2001. The contribution of two ears to the perception of vertical angle in sagittal planes. *The Journal of the Acoustical Society of America* 109 (4): 1596–1603. <https://doi.org/10.1121/1.1352084>.
- Morimoto, M., and H. Aokata. 1984. Localization cues of sound sources in the upper hemisphere. *Journal of the Acoustical Society of Japan* 5 (3): 165–173. <https://doi.org/10.1250/ast.5.165>.
- Musicant, A.D., and R.A. Butler. 1984. The influence of pinnae-based spectral cues on sound localization. *Journal of the Acoustical Society of America* 75 (4): 1195–1200. <https://doi.org/10.1121/1.390770>.
- Pastore, M.T., and W.A. Yost. 2017. Sound source localization as a multisensory process: The Wallach azimuth illusion. *The Journal of the Acoustical Society of America* 141 (5): 3635–3635.
- Perrett, S., and W. Noble. 1997a. The contribution of head motion cues to localization of low-pass noise. *Perception and Psychophysics* 59 (7): 1018–1026. <https://doi.org/10.3758/BF03205517>.
- Perrett, S., and W. Noble. 1997b. The contribution of head motion cues to localization of low-pass noise. *Perception and Psychophysics* 59 (7): 1018–1026.
- Pick, H.L., D.H. Warren, and J.C. Hay. 1969. Sensory conflict in judgments of spatial direction. *Perception and Psychophysics* 6 (4): 203–205.

- Pierce, A. 1901. *Studies in Auditory and Visual Space Perception*. New York: Longmans, Green, and Co.
- Platt, B.B., and D.H. Warren. 1972. Auditory localization: The importance of eye movements and a textured visual environment. *Perception and Psychophysics* 12 (2B): 245–248.
- Radeau, M., and P. Bertelson. 1974. The after-effects of ventriloquism. *The Quarterly Journal of Experimental Psychology* 26 (1): 63–71. <https://doi.org/10.1080/14640747408400388>.
- Radeau, M., and P. Bertelson. 1977. Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. *Perception and Psychophysics* 22 (2): 137–146. <https://doi.org/10.3758/BF03198746>.
- Radeau, M., and P. Bertelson. 1987. Auditory-visual interaction and the timing of inputs. Thomas (1941) revisited. *Psychological Research* 49 (1): 17–22.
- Rayleigh, L. 1876. On our perception of the direction of a source of sound. In *Proceedings of the Musical Association*, vol. 2, 75–84.
- Recanzone, G.H. 2009. Interactions of auditory and visual stimuli in space and time. *Hearing Research* 258 (1–2): 89–99. <https://doi.org/10.1016/j.heares.2009.04.009>.
- Searle, C.L. 1973. Cues required for externalization and vertical localization. *The Journal of the Acoustical Society of America* 54: 308. <https://doi.org/10.1121/1.1978213>.
- Shelton, B.R., and C.L. Searle. 1980. The influence of vision on the absolute identification of sound-source position. *Perception and Psychophysics* 28 (6): 589–96.
- Sivia, D.S., and J.S. Skilling. 2006. *Data Analysis: A Bayesian Tutorial*, 2nd ed. New York: Oxford University Press.
- Slattery, W.H., and J.C. Middlebrooks. 1994. Monaural sound localization: acute versus chronic unilateral impairment. *Hearing Research* 75 (1): 38–46.
- Slutsky, D.A., and G.H. Recanzone. 2001. Temporal and spatial dependency of the ventriloquism effect. *Neuroreport* 12 (1): 7–10.
- Solman, G.J., T. Foulsham, and A. Kingstone. 2017. Eye and head movements are complementary in visual selection. *Royal Society Open Science* 4 (1): 160569. <https://doi.org/10.1098/rsos.160569>.
- Spence, C., and J. Driver. 2000. Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. *Neuroreport* 11 (9): 2057–2061.
- Stein, B.E., and M.A. Meredith. 1990. Multisensory integration. Neural and behavioral solutions for dealing with stimuli from different sensory modalities. *Annals of the New York Academy of Sciences* 608: 51–70.
- Steinman, S.B., and R.P. Garzia. 2000. *Foundations of Binocular Vision: A Clinical perspective*, 2–5. McGraw-Hill Professional
- Stensola, T., and E.I. Moser. 2016. Grid cells and spatial maps in entorhinal cortex and hippocampus. In *Micro-, Meso- and Macro-Dynamics of the Brain. Research and Perspectives in Neurosciences*, ed. G. Buzsáki, and Y. Christen, 59–80. Cham: Springer. <https://doi.org/10.1007/978-3-319-28802-4>.
- Thompson, S.P. 1878. On binaural audition. *Philosophical Magazine* 2 (6): 383–391.
- Thurlow, W.R., and C.E. Jack. 1973. Certain determinants of the ‘ventriloquism effect’. *Perceptual and Motor Skills* 36 (3): 1171–1184. <https://doi.org/10.2466/pms.1973.36.3c.1171>.
- Thurlow, W.R., and T.P. Kerr. 1970. Effect of a moving visual environment on localization of sound. *The American Journal of Psychology* 83 (1): 112–118.
- Van Opstal, A.J. 2016. *The Auditory System and Human Sound-Localization Behavior*, 1st ed, 436. Amsterdam: Academic Press.
- Van Wanrooij, M.M., and A.J. Van Opstal. 2004. Contribution of head shadow and pinna cues to chronic monaural sound localization. *Journal of Neuroscience* 24 (17): 4163–4171. <https://doi.org/10.1523/JNEUROSCI.0048-04.2004>.
- Wallace, M.T., R. Ramachandran, and B.E. Stein. 2004. A revised view of sensory cortical parcellation. *Proceedings of the National Academy of Sciences* 101 (7): 2167–2172. <https://doi.org/10.1073/pnas.0305697101>.
- Wallach, H. 1938. Über die Wahrnehmung der Schallrichtung (On the perception of sound direction). *Psychologische Forschung* 22 (3–4): 238–266.

- Wallach, H. 1939. On sound localization. *The Journal of the Acoustical Society of America* 10 (4): 270–274.
- Wallach, H. 1940. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology* 27 (4): 339–368.
- Warren, D.H. 1970. Intermodality interactions in spatial localization. *Cognitive Psychology* 1 (2): 114–133. [https://doi.org/10.1016/0010-0285\(70\)90008-3](https://doi.org/10.1016/0010-0285(70)90008-3).
- Welch, R.B., and D.H. Warren. 1980. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin* 88 (3): 638.
- Wenzel, E.M., M. Arruda, D.J. Kistler, and F.L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94 (1): 111–123.
- Wightman, F.L., and D.J. Kistler. 1997. Monaural sound localization revisited. *The Journal of the Acoustical Society of America* 101 (2): 1050–63. <https://doi.org/10.1121/1.418029>.
- Wightman, F.L., and D.J. Kistler. 1999. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America* 105 (5): 2841–53. <https://doi.org/10.1121/1.426899>.
- Woodworth, R., and H. Schlosberg. 1938. *Experimental Psychology*. New York: Henry Holt and Company.
- Yost, W.A. 1981. Lateral position of sinusoids presented with interaural intensive and temporal differences. *The Journal of the Acoustical Society of America* 70 (2): 397–409. <https://doi.org/10.1121/1.386775>.
- Yost, W.A. 2016. Sound source localization identification accuracy: Level and duration dependencies. *The Journal of the Acoustical Society of America* 140 (1): EL14–EL19. <https://doi.org/10.1121/1.4898045>.
- Yost, W.A. 2017a. History of sound source localization: 1850–1950. *The Journal of the Acoustical Society of America* 30: 1–15. <https://doi.org/10.1121/2.0000529>.
- Yost, W.A. 2017b. Sound source localization identification accuracy: Envelope dependencies. *The Journal of the Acoustical Society of America* 142 (1): 173–185. <https://doi.org/10.1121/1.4990656>.
- Yost, W.A., M.T. Pastore, and K.R. Pulling. 2019. Sound source localization as a multisystem process: the Wallach azimuth illusion. *The Journal of the Acoustical Society of America* 146 (1): 382–398.
- Yost, W.A., M.T. Pastore, and M.F. Dorman. 2020. Sound source localization is a multisystem process. *Acoustical Science and Technology* 41 (1).
- Yost, W.A., X. Zhong, and A. Najam. 2015. Judging sound rotation when listeners and sounds rotate: Sound source localization is a multisystem process. *The Journal of the Acoustical Society of America* 138 (5): 3293–3310. <https://doi.org/10.1121/1.4920001>.
- Zahorik, P. 2002. Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America* 112 (5): 2110. <https://doi.org/10.1121/1.1506692>.
- Zahorik, P., P. Bangayan, V. Sundareswaran, K. Wang, and C. Tam. 2006. Perceptual recalibration in human sound localization: Learning to remediate front-back reversals. *The Journal of the Acoustical Society of America* 120 (1): 343–359. <https://doi.org/10.1121/1.2208429>.
- Zhong, X., L. Sun, and W.A. Yost. 2016. Active binaural localization of multiple sound sources. *Robotics and Autonomous Systems* 85: 83–92. <https://doi.org/10.1016/j.robot.2016.07.008>.
- Zwiers, M.P., and A.J. van Opstal. 2001. A spatial hearing deficit in early-blind humans. *Journal of Neuroscience* 21 (9): RC142.

Spatial Soundscape Superposition and Multimodal Interaction



Michael Cohen and William L. Martens

Abstract Contemporary listeners are exposed to overlaid cacophonies of sonic sources, both intentional and incidental. Such soundscape superposition can be usefully characterized by where such combination actually occurs: in the air, at the ears of listeners, in the auditory imagery subjectively evoked by such events, or in whatever audio equipment is used to mix, transmit, and display such signals. This chapter regards superposition of spatial soundscapes: physically, perceptually, and procedurally. *Physical* (acoustic) superposition describes such aspects as configuration of personal sound transducers, panning among multiple sources, speaker arrays, and the modern challenge of how to integrate and exploit mobile devices and “smart speakers.” *Perceptual* (subjective and psychological) superposition describes such aspects as binaural image formation, auditory objects and spaces, and multimodal sensory interpretation. *Procedural* (logical and cognitive) superposition describes higher-level relaxation of insistence upon literal auralization, leveraging idiom and convention to enhance practical expressiveness, metaphorical mappings between real objects and virtual position such as separation of direction and distance; range-compression and -indifference; layering of soundscapes; audio windowing, narrow-casting, and multipresence as strategies for managing privacy; and mixed reality deployments.

1 Introduction: Stages of Composition

Auditory displays are broadly and richly embedded in modern life. We are positively assailed by communication sounds, competing with each other for attention. Spatial soundscape superposition can be usefully characterized by where the combination

M. Cohen (✉)
Spatial Media Group, Computer Arts Laboratory, University of Aizu,
Aizu-Wakamatsu, Fukushima 965-8580, Japan
e-mail: mcohen@u-aizu.ac.jp

W. L. Martens
Discipline of Physiology, School of Medical Sciences, Faculty of Medicine and Health,
University of Sydney, Sydney, NSW 2006, Australia

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_13

Table 1 Spatial soundscape superposition

Stage	Domain	Realm	Practice	Considerations
<i>Sound</i>	Acoustics	<i>Physics</i>	<i>Transmission:</i> air mixing plus bone-conduction	personal and public transducers, panning, speaker arrays, smart speakers, mobile-ambient interfaces
Transduction	Biophysics, biochemistry	Physiology	Cochlear implants	Critical bands and ERBs, auditory or loudness recruitment
<i>Sensation</i>	Psychology, psychoacoustics	<i>Perception:</i> sensorineural processes	<i>Apprehension:</i> subjective composition, vection	Auditory objects, binaural imagery, multimodal stimulation
<i>Signals</i>	Cognition	<i>Procedure:</i> central auditory process	<i>Interpretation:</i> logical convention, metaphorical mapping and mixing, culture and semiotics	Parameterized directionalization and spatialization, layering and audio windowing, mental models, practical interpretation

actually occurs. As anticipated by Table 1, this chapter regards superposition of spatial soundscapes: physically, perceptually, and procedurally—sound, sensation, and signal, following Hartmann’s titular description (1999) of the auditory process (albeit in rotated order).

2 Physical Superposition (Air Mixing): Sound

Normal circumstances combine sound in air, as when ordinary sources such as voices naturally add. The air acts as a linear mixer, superposing respective pressure disturbances. Modern instances of such superposition involve electroacoustics, using speakers to display some organized diffusion, such as sound distribution and panning. Physical combination leverages installed speakers as well as mobile devices such as cell phones, laptop computers, and smart speakers. Stereo speakers, sound bars, and home theater systems are common installations. In environments such as automobiles, ordinary loudspeakers can be replaced with novel actuator systems, such as distributed mode actuators (DMAs), distributed mode loudspeakers (DMLs), and multiactuator panels (MAPs). For example, the Ac2ated Sound (<https://continental-automotive.com/en-gl/Passenger-Cars/Information-Management/Multimedia-Systems/Ac2ated-Sound>) system attaches transducing drivers to car interior elements, using the pillars and dashboard for high- and mid-frequency reproduction, and large components—such as the ceiling, back covers of seats, and rear shelf—for low frequencies. More specialized spaces have super-directional (sound beam) loudspeakers

and phased arrays. Some speakers, such as the Nexo CDD (configurable directivity device) (<https://nexo-sa.com/systems/geo-s12/technology/>), even feature adjustable directivity.

Directly connecting speakers to microphones, as in simple channel-based telepresence installations—dating back to the Théâtrophone exhibited at the Paris Electrical Exhibition in 1881, and formalized by Alan Blumlein (https://en.wikipedia.org/wiki/Alan_Blumlein) in the 1930s—can recreate sound fields. More active manipulations process audio streams by time delay and filtering, which can be implemented in digital signal processing (DSP) systems via recursive delay-and-add networks, or as time-domain convolution or equivalent frequency-domain multiplication.

2.1 Speaker Arrays

Besides theatrical spatial sound systems and the *sui generis* Audium (<http://www.audium.org>)—shown in Fig. 1—various institutions maintain polyphonic media art centers and concert halls, including such high-density loudspeaker arrays (HDLAs) (Lyon 2016, 2017) as the AlloSphere in Santa Barbara, California (<http://www.allosphere.ucsb.edu>), the BEAST (Birmingham ElectroAcoustic Sound Theatre) (<http://www.beast.bham.ac.uk>) project (Birmingham, UK), CCRMA at Stanford University (<https://ccrma.stanford.edu>) Stanford University, Palo Alto, California; Espace de PROJECTION (“Espro”) (<http://web4.ircam.fr/1039.html?&L=1>) at IRCAM (Paris), “The Cube” (<http://icat.vt.edu/studios.html>) at Virginia Tech’s Institute for Creativity, Arts, and Technology (ICAT), and the Spatial Sound Institute, Budapest, Hungary (<https://spatialsoundinstitute.com>). Annual festivals highlight multichannel sound, including Berlin’s Club Transmediale (<https://transmediale.de>), Edmonton’s Sea of Sound (<http://www.beams.ca/SeaofSound.htm>), and Ontario’s New Adventures in Sound Art (<http://naisa.ca>).

Audio diffusers such as sound field renderers can pantophonically (horizontally) and periphonically (horizontally and vertically) distribute parallel inputs across speaker arrays using a mixer as a crossbar directionalizer. Such architecture scales up to arbitrary degrees of polyphony: multichannel songs, conference chat-spaces, and immersive soundscapes can be dynamically displayed via such controllers. For instance, a dynamic map interface, like that shown in Fig. 2a, can control distribution of multiple channels across a ring of speakers, like that in Fig. 2b, panning signals across an adjustable spread (or “aperture”) of speakers.

Spatial sound display is receptive to any number of modulations. Perception of situated sources includes impression not only of position and emission characteristics (relative location and orientation directivity), but also environmental effects, such as reflection, occlusion, obstruction, echo, and reverberation, as measured by such related metrics as Reverberation Time RT_{60} , Definition D_{50} , Clarity C_{80} , and interaural cross-correlation (IACC; <http://asastandards.org/Terms/interaural-cross-correlation/>).

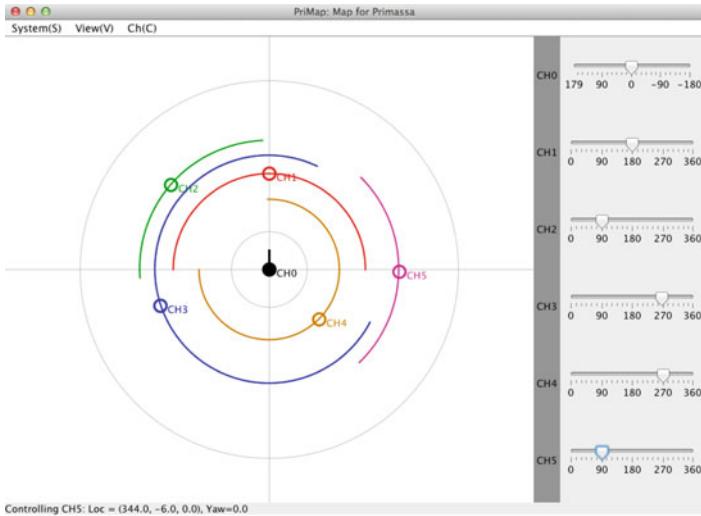


Fig. 1 The Audium (San Francisco)

Adjustment of the virtual position of sources can be independent of the underlying audio streams or somehow related. For simple example, a virtual musical source might encode harmony by moving around a space to signal chord progression (Herder and Cohen 2002) as a pedagogical tool. Dynamic gestures, auditory vectors comprising moving sources, can be used not only for spatial music but for “acoustic arrows,” animated sonic beacons for way-finding and multimodal displays, accompanying such verbal directions as “come hither and proceed thither.” Likewise, a simulated environment can be adjusted and parameterized by such variables as spatial dimensions, geometry, and liveness (absorption and diffusion material characteristics).

Besides articulated sound directionalization and spatialization, distributed display allows extended diffusion. Spatial extent can be suggested by multiple virtual sources and/or loudspeakers driven together, which resultant auditory (or apparent) source width (ASW) is interpretable as line, area, or volume sources. Much the same way that vibrato makes a note seem louder (Wolfe 2018), or aural exciters (which add high-order harmonic extensions to a signal) enhance conspicuity, wiggling a source or pulsating its size can make it “shimmer” to stand out. To draw attention to a virtual source, an aware agent (a software component that monitors, confirms, and sharpens user focus) can modulate various aspects, including perturbing its position, and dilating and contracting it (adjusting its spatial volume). Such highlighting can push a track to prominence in a mix, like a musical warble, trill or quaver.

Perceptual rivalry—such as contradictory IID (interaural intensity difference, or head shadowing, a.k.a. ILD, interaural level difference) and ITD (interaural time delay)



(a) ASw pantophony: visualization and control interface for sink azimuth and concentric auditory source widths (ASWs). Arc angles correspond to diffusion aperture across an annularly arranged (circumferential) speaker array. (Map and diffuser developed by Yoshiyuki Yokomatsu.)



“Come out with your hands up—you’re surrounded by men with megaphones.”

(b) Megaphonic pantophony. (©2020 The New Yorker Collection from cartoonbank.com. All rights reserved.)

Fig. 2 Pantophonic perimeter

cues—leads to diffuse source imagery, which a listener describes as a “fuzzy” region. Such localization blur can be thought of as the resolution of spatial hearing.

Because of the horizontal arrangement of our ears, paralleling the gravity-oriented arrangement of our limbs and eyes, IID and ITD panning affect virtual source azimuthal direction (but not elevation or range); it is easier to create auditory “*bokeh*” (out-of-focus blurring) laterally than vertically or longitudinally.

2.2 *Panning*

A panoramic potentiometer (or “pan-pot”) can control distribution of audio power across multiple speakers. To avoid panned signal coherence from disturbing broadened display (via such artifacts as the precedence or Haas effect), image dispersion and signal decorrelation (via such DSP techniques as all-pass filtering, vibrato, and chorusing) can be used to scramble the phase of frequency components (Kendall 2010). DBAP: Distance-based Amplitude Panning (Lossius et al. 2009), MIAP: Manifold-Interface Amplitude Panning (Seldess 2014), VBAP: Vector Base Amplitude Panning (<http://legacy.spa.aalto.fi/research/cat/vbap/>) (Pulkki 1997) and DirAC: Directional Audio Coding (Pulkki et al. 2011, <http://legacy.spa.aalto.fi/research/cat/DirAC/>) can be considered generalizations of pan-pots. Phased array and beam-forming soundfield synthesis (SFS) techniques—such as wavefield (or wavefront) synthesis (WFS) and boundary surface control (BoSC) or boundary element methods (BEMs)—which modulate not only gain but also the delay and spectra of multiple signals, require active signal-processing, more aggressive DSP than just amplitude modulation and frequency filtering.

2.3 *Personal Listening Systems, PSAPs, Hearables, Hearing Aids, and XR (Extended Reality—AR: Augmented Reality, MR: Mixed Reality, and VR: Virtual Reality)*

The contemporary panoply of personal listening devices is surveyed and summarized by Fig. 3. Besides ordinary speakers, personal sound amplification products (PSAPs) and hearing aids are increasingly popular and important, performing vigorous DSP, including directional capture, filtering and active noise control (or cancellation, ANC), ameliorating sensorineural and conductive hearing loss such as presbycusis, age-related hearing loss, as well as compensating for loudness recruitment, rapid increase with amplitude of perceived loudness. By detecting characteristics of an environment, audio processing can automatically change parameters to accommodate various situations (conversation, restaurant, TV, cinema, driving, telephone, concert, etc.). Equalization can be done monaurally using ipsilaterally embedded processors, but binaural processing can be performed on a smartphone, including “diminished

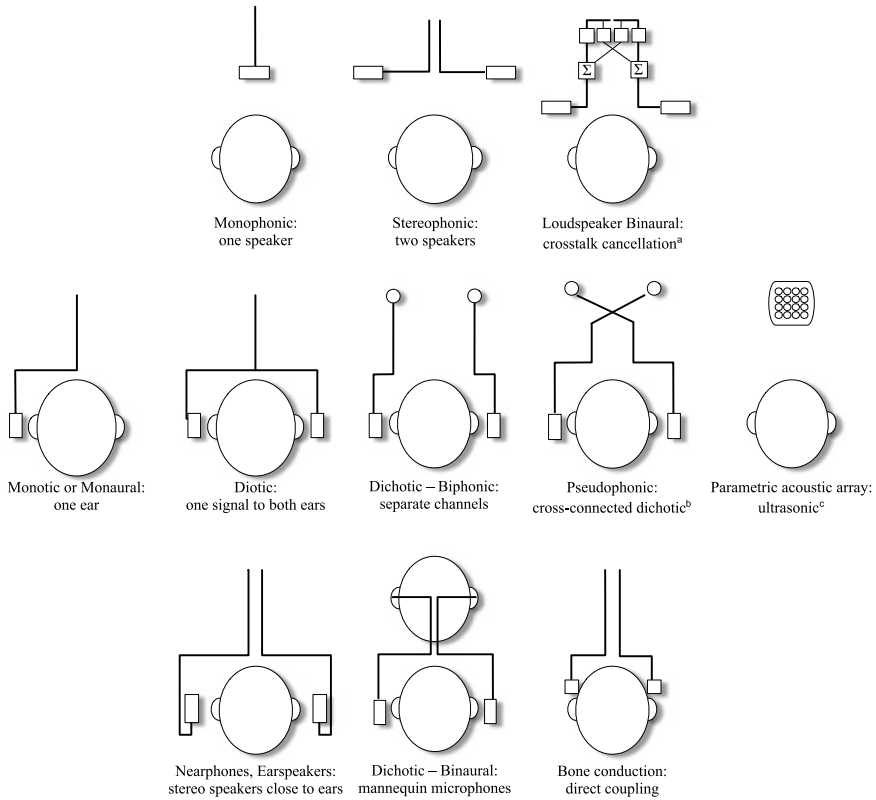


Fig. 3 Personal sound displays: A variety of form factors for personal audition, arranged in order of intimacy. The drivers, symbolized by rectangles, may be wireless, such as Bluetooth earbuds and headsets. (Extended by Cohen 2016 from Streicher and Everest 2006 and Marui and Martens 2006.) **a** By preconditioning stereo signals, speaker crosstalk can be controlled and significantly cancelled or compensated for (cross-talk cancellation, CTC). A special implementation of this technique is called “transaural” (Bauck and Cooper 1996; Choueiri 2018). **b** Pseudophonic arrangements allow dramatic demonstration of the importance of active, head-turning directionalization, as front-back and up-down disambiguations are subverted, even if a subject can see the source (Martens et al. 2011). **c** Ultrasonic displays (a.k.a. parametric loudspeakers), such as that described by Ochiai et al. (2017), represent a special case: inaudible ultrasonic signals demodulate in the air, so the audible source is the air itself, not the driver. Somewhat similarly, some new displays exploit the photoacoustic effect (Sullenberger et al. 2019), by which sound is formed as a result of material absorbing light, such as a laser beam

reality” off-axis rejection for focused hearing. Such architecture also supports disintermediation, eliminating unnecessary dataflow stages: hearing assistance transmission systems using induction looping (telecoils) or FM radio can be replaced by beaming stereo streams directly to earphones, avoiding cumbersome reconstruction, transduction by external speakers, and recapture by hearing aid microphones before resynthesis by in-ear drivers. Hearing aids feature “superhearing” processing

(hyperacuity: hypersensitivity and hyperselectivity) informed by auditory scene analysis (Bregman 1990) and deep learning (Wang 2017), enhancing sound segregation and isolating speech. “Human hacking”—bionic augmentation such as prostheses, cochlear implants, and bone-anchored hearing aids—invites extended auditory displays.

Contemporary personal audition systems also include virtual reality (VR) and augmented reality (AR) auditory displays, typically using head-mounted displays (HMDs). VR and AR are generally considered mixed reality (MR), so the abstraction of all of these in current parlance is XR, where the ‘X’ stands not only for “extended” but also for “Augmented,” “Mixed,” and “Virtual” (as at the end of this subsection’s title). XR can be applied to visualization of sound fields by overlaying visual intensity indication upon actual acoustic spaces (Inoue et al. 2017), but more relevantly and importantly, it can leverage environmental or ambient resources for richer soundscapes. This subject is revisited below in Sect. 2.5.

Some exotic headphones highlight innovative capability, calibrating for anatomy or featuring head-tracking and multidriver arrays to emulate directional sources. For example, the Sennheiser Ambeo headphones (<https://en-us.sennheiser.com/in-ear-headphones-3d-audio-ambeo-smart-headset>) feature ANC, “hear-through” acoustic transparency, and binaural recording. Bose Frames (https://www.bose.com/en_us/products/frames.html) sunglasses have earstem-embedded, personal back-firing speakers, a microphone for voice control and conferencing, Bluetooth connectivity, and head-tracking for AR applications such as audio tour guides. The Panasonic Wear Space (<https://panasonic.net/design/flf/works/wear-space/>) features ANC wireless headphones extended with head-wrapping fabric, enhancing concentration by blocking noise and peripheral visual distractions. Nura headphones (<https://www.nuraphone.com>) have a circumaural body combined with earbuds; its set-up calibration analyzes otoacoustic emissions (OAEs), weak sound generated by the cochlea, to adjust equalization; and tactile bass is delivered through “immersion mode” ear-cup drivers. Sony 360 Reality Audio (<https://www.sony.com/electronics/360-reality-audio>) headphones, calibrated by probe microphones, are part of a larger system dedicated to flexible display of 3D audio. The “Aware” headphone (<http://www.unitedsciences.com/the-aware-kickstart-the-hearable-revolution/>) or “hearable” (<https://www.everydayhearing.com/hearing-technology/articles/hearables/>) has integrated EEG (electroencephalography) sensors, allowing estimation of a wearer’s mental state (as reviewed below in Sect. 5.1).

2.4 Panic in the Anech: Extending Live Direct Sound with Environmental Indirect Sound

Another type of physical superposition does not usually employ binaural technology, but becomes very interesting when it does. If a violin is played under anechoic conditions, or captured in a non-reverberant practice room, the performer will typically dislike the unnatural character of the sound—that is, “panic in the anech[oi

chamber].” A commonplace non-binaural solution is to submit the ‘dry’ input source to reverberation processing and loudspeaker reproduction to create the more musically familiar ‘wet’ sound signal, so that the performer can hear the sound of their violin in a manner more typical of an acoustically live performance space. Now imagine the binaural counterpart to this, where the direct sound of the violin is captured by a closely-placed, instrument-mounted (“spot”) microphone, and this signal is processed for binaural display such that the indirect sound of a reverberant space responding to the instrumental sound is realistically reproduced via ear-speakers (drivers positioned near but not on the auricles, without circumaural cushions or contact with the pinnae), deployed to allow direct sound from the violin to enter the ears without interference. The performer hears the direct sound from the violin as usual, but with plausibly realistic binaural information in the reproduced indirect sound superposed upon it. This can be valuable for a performer during rehearsal, as the enriched reproduction can mimic the acoustics of the performance space for which they would like to be prepared.

Similarly, when speaking or singing, one’s voice returns to one’s ears with information about the room and its interaction with the voice, yielding an impression of the space. The room acoustical contribution to the sound of one’s voice can be represented via the Oral-Binaural Room Impulse Response (OBRIR), so that self-generated ‘direct’ sounds can be combined in the air (i.e., air-mixed, including the ever-present, bone-conducted, vocal sound: the “human sidetone”) with environmental ‘indirect’ sound that has been electroacoustically introduced (Cabrera et al. 2009). In one such deployment, indirect sound associated with a sound source was reproduced via a pair of ear-speakers, so that binaurally recreated indirect sound could be added to unobstructed ‘live’ sound propagating directly from mouth to ear.

A converse arrangement that also relies upon acoustically transparent ear-speakers is that which might be used to superimpose virtual sound sources upon ‘live’ environmental sound so as to minimize interference of the ear-speakers with natural spatial hearing. For example, in augmented spatial auditory displays providing navigational aid to visually challenged users, minimized interference from a binaural auditory display system is required, since navigation by the blind can be enabled through use of available sonic information, often with refined skills using sound alone. Removing this “open-ear” channel by covering the pinna or plugging the ear canal with insert earphones would disable a needed sensory system, causing drastic reduction in the considerable acuity such users exhibit with their own natural spatial hearing for navigation.

Clear directional imagery was demonstrated for speech signals using such an open-ear binaural superposition system, developed with the commercially available “TOPlay” Open Guided Sound (OGS) earphones (Pereira and Martens 2018; <http://www.toplay-ogs.com>). Speech signal localization performance using OGS earphones, featuring so-called “TrueOpen” technology to deliver sound directly to the ear-canal entrance with minimal obstruction of the pinnae, was comparable to that assessed using a 196-channel loudspeaker array. Additional “mobile-ambient” systems are discussed in the following subsection.

2.5 Mobile-Ambient Systems: Combination of Personal and Public Displays

Table 2 shows a variety of audio and visual output devices, ordered by intimacy. In analogy to laptop and desktop computing, “eartop” and “eyetop” form factors describe closely attached personal displays. Eartop transducers featuring sound displays for individuals can be integrated with public loudspeaker systems. Even closed-back or circumaural headphones are not completely acoustically opaque, leaking sound in both directions. That is, ambient speakers can be used to complement headphone-displayed soundscapes.

In situations where public and private resources are both available, combinations can leverage advantages of each. As suggested by Fig. 4, hybrid configurations will emerge, such as loudspeaker arrays in conjunction with eartop displays (Satongar et al. 2015) and arrangements of mobile phone speakers. A cinema could feature individual binaural channels, like those served by SoundFi (<http://soundfi.me>), as well as the theatrical multichannel system, for personalized auditory display, including localized dialog and multilingual narration. Bass management might route low-frequency effects (LFES) to shared subwoofers whilst sending higher frequency bands to per-

Table 2 Audio and visual displays along private ↔ public continua

Proxemic context	Architecture	Display	
		Audio	Visual
Intimate, Personal, Private	Headset, XR, wearable computer	<i>Eartop</i> (earwear), headphone, earbud, earphone, hearing aid, PSAP, hearable, in-ear monitor, bone-conduction (cheekbone, neckband, collarbone, ...)	<i>Eyetop</i> (eyewear), HWD (head-worn display), HMD (head-mounted display)
Individual	Chair	Smartphone, nearphone, ear-speaker, “sound shower” isolation directional display	Smartphone, tablet, <i>laptop</i> display, <i>desktop</i> monitor
Interpersonal	Couch or bench	Loudspeaker (e.g., stereo dipole, transaural TM)	HDTV, “fishtank VR”
Multipersonal, Familiar	Home theater, vehicle, spatially immersive display (e.g., Cave, TM Cabin)	Surround sound, soundbar, ITU 5.1, 7.1.4, NHK 22.2, etc.	Projection, 4K, 8K
Social	Club, theater	Speaker array (e.g., VBAP, DirAC, DBAP, WFS)	Large-screen display (e.g., IMAX)
Public	Stadium, concert arena	Public address system, (additional sound reinforcement, with delay towers for distant listeners), siren, klaxon	Multiple screens (additional image display to reach distant viewers)

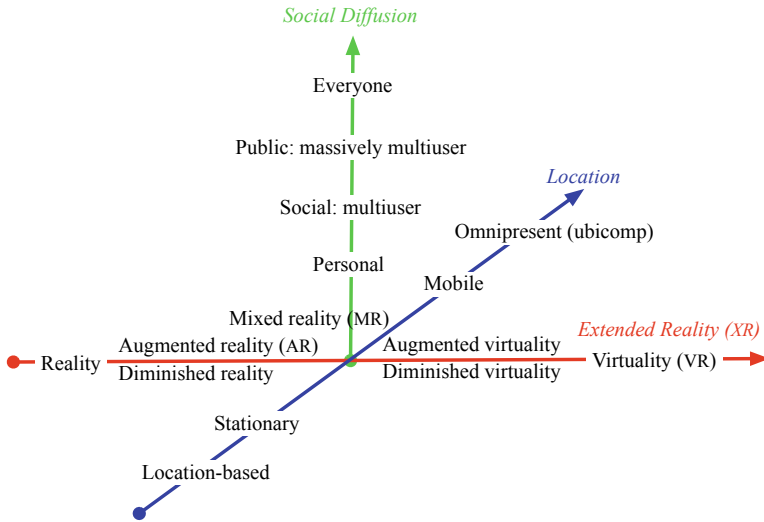


Fig. 4 Extended reality (XR), location, and social diffusion taxonomy—The horizontal Extended Reality (XR) axis is the original AR–VR continuum (Milgram and Coquhoun, Jr. 1999); **Location** (longitudinal axis) refers to where such XR systems are used; **Social Diffusion** (vertical axis) refers to degree of concurrent usage. Adapted and extended from (Broll et al. 2008)

sonal transducers. The dichotomy between mobile computing and site-specific LBS (location-based services) is resolved with “mobile-ambient” transmedial interfaces that span both personal, mobile devices and public, shared resources (Cohen 2016).

2.6 Implications: IoT and UbiComp

Global popularity of mobile computing creates opportunities for new kinds of computer-human interaction, including democratized control and distributed display. For instance, even technophobes uncomfortable with personal computers can enjoy rich interaction with smartphones. The social diffusion of wireless devices has been paralleled by a separate development of networked appliances: internet of things (“IoT”), ubicomp (ubiquitous computing), and pervasive computing. Sensors and displays will increasingly find their way into everyday circumstances, allowing exploitation by roomware media managers, software for smart buildings.

In computer graphics, “projection mapping” refers to adjusting presentation for display on irregular surfaces, preconditioning contents to anticipate a physical space into which a scene is projected. Auditorily, flexible sound renderers encourage such display context-sensitivity. A simple example is a loudspeaker crossover circuit, which frequency-band filtering matches spectral responses of a multidriver speaker. A more novel example is an opportunistic mixer that routes channels among available

Table 3 Saturated: distributed and pervasive, continuous and networked, transparent or invisible—spatial hierarchy of ubicomp or ambient intimacy

- Smart spaces, smart cities, urban (or street) computing
- Cooperative or intelligent buildings and smart homes
- Roomware and reactive rooms
- Spatially immersive displays
- Information furniture
- Networked appliances, smart displays
- Handheld, mobile, nomadic, portable, and wireless (unplugged) devices
- Wearable computers, smart watches, smart glasses, hearables, XR HMDs
- Computational clothing (smart clothes), hearing aids, PSAPs

resources, discovered and managed by smart homes, intelligent building controllers, “urban (or street) computing,” and “smart city” infrastructure. As outlined by Table 3, displays should collaborate across all scales. (These ideas are revisited below in Sects. 5.1 and 5.3.)

3 Perceptual Superposition (Subjective Compositing): Sensation

Whereas the previous section of this chapter dealt with the great variety of physical soundscape superposition to which listeners are exposed, this section addresses perceptual experiences associated with such exposure. The treatment recognizes the complexity of binaural image formation when listeners move relative to sound reproduction systems whilst simultaneously receiving sensory input through multiple modalities, including not only auditory, but also visual and vestibular systems (Martens and Cohen 2020).

Perceptual superposition depends, of course, upon binaural stimuli presented via physical superposition (appearing as afferent signals), but spatial hearing also depends on observers being aware of their own motion in the world (perhaps through efferent signals associated with motor commands, but also through cognitive factors that exert top-down influences on operations such as binaural image formation).

Because perception can be influenced as much by cognitive factors as by stimulus parameters, purely bottom-up (signal-driven, or afferent) models of spatial perception sometimes yield poor predictions of human experience. This is particularly evident in results of studies that include listening conditions allowing listener movement, such as listening while walking (Martens et al. 2011). Although it is difficult to experimentally determine the role of binaural cognition, as scientific studies focus predominantly upon overt behavior, it is reasonable to suppose that cognitive factors (based, for example, upon expectations) operate during listener movement by disam-

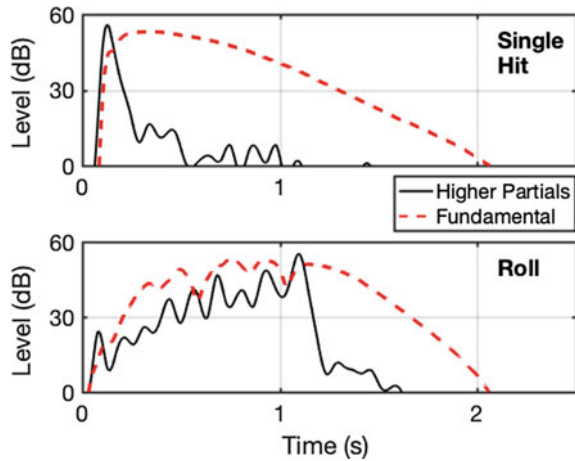
biguating raw sensations through implicit hypothesis testing, such as that associated with “symbol processing” (Blauert et al. 2013; Blauert 2017). Auditory scenes are mentally constructed in the context of potentially abstract thoughts and concepts associated with procedural superposition, which is taken up in the next section of this chapter. Before delving into that topic, the process of binaural image formation shall be discussed, and the complexity of this process, appearing superficially simple, will be revealed.

3.1 *Binaural Image Formation: Perceptual Fusion (Integrated Superposition) and Fission (Segregation)*

Binaural image formation is the process by which acoustic events to which listeners are exposed lead to the experience of associated auditory events. These auditory events comprise auditory objects that are heard to be located in auditory spaces. While this seems straightforward enough, the process of auditory image formation is neither simple nor well understood. Indeed, there is not always a one-to-one relationship between acoustic events and auditory events. Single acoustic events may give rise to multiple auditory events: perceptual fission (segregation) has occurred. Multiple acoustic events may give rise to only one auditory event: perceptual fusion (integrated superposition) of incoming energy into a coherent entity has occurred. Superposition of sonic events that are presented with the intention of creating an integrated unitary percept will not necessarily be successful, so principles of fusion and fission are examined here. Under typical binaural listening conditions, when the sounds of an external acoustic event impinge upon ears of a human listener, an auditory image of a sounding object typically results. This auditory image may or may not be heard as externalized, i.e., heard as occurring outside the listener’s head. If externalized, the auditory image may be described as an *auditory object*, a mental representation associated with an acoustic event resulting from perceptual fusion of the incoming sound energy into a single, coherent entity. In discussion of binaural image formation, this distinction between acoustical and auditory events should be clearly defined: sounding objects associated with acoustic events have *actual* positions in the physical space surrounding the listener; associated auditory objects have *apparent* positions in auditory space, a mentally constructed space in which auditory events can occur. Acoustic events that occur in reverberant environments are usually heard as occurring outside a listener’s head (i.e., as externalized auditory objects), and yet it is important to recognize these auditory objects as mental projections into psychological constructions of those reverberant environments as they are perceived.

In the context of this discussion on soundscape superposition, understanding principles underlying binaural image formation is key to linking physical superposition and perceptual superposition. This is not a new idea. Plenge (1974) proposed that a sound stimulus should form a coherent auditory image if and only if natural processes

Fig. 5 Examples of temporal envelopes of frequency components for two types of marimba performance, where the dashed curves show the envelope for the fundamental frequency and the solid curves show the sum of the higher-frequency overtones



of spatial hearing are engaged. His model stressed that sound localization has as its first condition...

[...] the ability, learned in early childhood, to classify [auditory] events as sound events. This ability may comprise, besides the perception of direction and distance, the ontogenetic earlier fusion of the information coming through both ears into one general acoustic image.

In free-field sound localization research, asking a listener to report the location of a sound stimulus is reasonable, even when the sound stimulus is as simple as a gated sinusoid. But when a listener uses headphones, such simple stimuli are often heard as within the listener's head ("IHL": inside-the-head-locatedness (Wenzel et al. 2018)), under which conditions Plenge (1974) would term the task *lateralization* rather than *localization*. Even when broadband binaural stimuli are employed, there is no guarantee of externalization and coherent auditory imagery (Toole 1969).

Consider the auditory imagery associated with the binaural presentation of a musical note played on a marimba. Even when a high-quality microphone captures a dry but realistic sounding marimba performance, and then that signal is transformed for headphone presentation through a listener's own measured head-related or anatomical transfer functions (HRTFs or ATFs), the fundamental frequency component of the marimba note typically segregates spatially from the higher-frequency partials of the note which decay more rapidly (and correspond to the brief "strike tone," rather than the more slowly decaying resonance corresponding to the nominal pitch (Perrott et al. 1987)).

For the "single hit" marimba performance shown in the upper panel of Fig. 5, it is easy to see how there might be segregation based on the difference in the temporal envelope of the fundamental frequency component versus that of the higher-frequency partials, which are summed to produce the single solid curve. If, however, a series of rapid marimba notes is performed as in the "roll" performance in the lower panel, listeners have the opportunity to rotate their heads while listening. The two temporal envelopes, while not strictly correlated, nonetheless rise and fall

together, so that coordinated lateral shifts in the tone's fundamental and higher partials accompany head-turning or "idling" postural sway. Whether listeners use their natural head acoustics, or use a headphone-based binaural display incorporating active head-tracking, there is an increased likelihood of perceptual fusion of all these frequency components in this dynamic case. Then, if the presentation includes an effective (i.e., spatially realistic) binaural simulation of indirect sound, the binaural image of the marimba tones will likely be heard as both unified and externalized. It is tempting to propose that a Gestalt principle could be operating, where the fundamental frequency that normally segregates from the strike tone of each note might be integrated based upon the 'common fate' of all partials as they shift in lateral angle in response to head-turning.

Whereas in free-field conditions it would be reasonable to elicit a report of the direction and distance of the marimba as a sounding object in physical space, without head-tracking, headphone presentation of a spatially static and dry marimba tone creates a complex percept that cannot be assigned a single direction or position in space. For many years, much of the spatial hearing literature considering headphone presentation has obscured this issue by using the term "localization judgments" to identify such estimates of the position of auditory objects.

Decades ago, Shaw (1982) argued for the importance of a distinction between performance in localizing sound objects and the ability to report the direction and distance of an auditory object experienced during headphone listening. He proposed that headphone studies of auditory spatial imagery be referred to as *space perception* rather than *sound localization*. If this sage advice had been heeded, considerable misunderstanding in the literature might have been avoided. Coupled with an emphasis on spatially static sources and listeners, many reported research results have contributed less to practical applications of binaural technology than desired. Philosophical underpinnings of the above issues are well addressed in a paper by Blauert (2012) that introduces into this discussion the concept of "Perceptionism." A perceptionist's approach to psychoacoustics is also a perspective on methods used in evaluating effectiveness of binaural technology, emphasizing methods that should benefit those engaged in optimizing spatial auditory display technology for real-world applications rather than artificial arrangements in research laboratories.

3.2 Moving Listeners: Dynamic Multimodal Sensory Integration

Much recent research regarding multimodal sensory integration in spatial hearing relates to the importance of voluntary motion in allowing listeners to understand changes in binaural stimuli coupled with changes in the orientation and position of those listeners (Pastore et al. 2020, this volume). Particularly telling in this regard are the results of studies using pseudophonic displays that swap signals between the left and right ears—as shown in this chapter's Fig. 3 and described by its caption b. For example, when listeners are fitted with pseudophonic displays that afford a "live"

interchange between left and right ear signals, and are then instructed to walk through an environment attempting to localize sources such as speech sounds, the naturally occurring head-motion-coupled variation in interaural directional cues dominates other localization cues (Martens et al. 2011). If, however, sources with emphasis on higher-frequency content are presented from stable “world-centric” positions, there is less dominance of head-motion-coupled changes in low-frequency interaural cues over spectral cues associated with the pinna. In fact, directional ambiguities can result from the cue conflict that results from such pseudophonic displays when broadband noise bursts are localized (Martens et al. 2013). However, when speech is the stimulus, continuous changes in orientation of the head during walking (such as head-turning) contribute to the creation of strong auditory illusions that are hard to suppress, even when the mouth of the talker is clearly visible. That so-called “Phantom Walker” study showed that when listeners with swapped left and right ear signals were asked to walk past a continuously viewed speech source emanating from a fixed spatial position, the source was heard to be moving through space at twice the listener’s rate, and arriving from a spatial region that was reversed with regard to all three spatial axes: left for right, front for back, and above for below. For example, despite having the stationary talker producing the speech sound in clear view as listeners walked toward that talker (where the “ventriloquism” effect might operate), the sound was invariably heard to be approaching from behind, and the voice of this illusory Phantom Walker overtook listeners as they passed by the physically stationary source. These head-coupled interaural cues are so strong that they defeat the contradictory “pinna-based” directional cues, as well as the visual cues (anchored on the actual talker).

Such observations have also been made in studies in which listeners were asked to turn their heads in a constrained fashion while dorsally located loudspeakers presented sources that shifted laterally across the rear hemifield, creating illusions of frontward incidence (Macpherson 2013), through a reversal of interaural cues accompanying head-turning. While these results replicate those of the classic study by Wallach (1940), a related, but possibly surprising result emerged when walking listeners rolled their heads while listening to speech sounds arriving from elevated loudspeakers in an analogous reversal of interaural cues accompanying head-rolling (Martens et al. 2011). Just as front-back reversals are associated with pseudophonic treatment during head-turning (Brimijoin and Akeroyd 2012), above-below reversals were shown to be associated with pseudophonic treatment during head-rolling (with cueing of source elevation depending on the resulting lateral shifts of source images). As have results of other related studies, Kawaura et al. (1991) suggest the dominance of dynamic interaural cues over spectral directional cues, at least for speech sounds containing energy mostly below 5 kHz. When sources containing more high-frequency energy are presented, presumably allowing pinna-based spectral cues greater influence on binaural image formation, the rate of these illusory reversals is greatly reduced (Martens et al. 2013).

To be clear, such head-motion-coupled directional cues do not require or depend upon gross listener movements. Indeed, even when listeners are asked to remain still during a sound localization task, they still move their heads by small but measurable

amounts (Wersényi and Wilson 2015), and they seem to move their heads just as much when engaged in natural listening activities, such as watching movies (Kim et al. 2013). Again, these recent studies of vestibular and other motion-based influences on binaural perception of auditory direction are preceded by important earlier studies. In introducing the topic of such non-acoustic influences on binaural perception, Lackner (1983) noted that studies of directional hearing conducted with a fixed head position and orientation clarify only part of the human capacity for spatial hearing:

Ordinarily a person is freely moving about and his head and trunk position vary both respect to each other and to external objects. Under these conditions the auditory cues at the ears from a stationary sound source change continuously. [...] In localizing an external sound source a person thus must monitor not only the auditory cues he receives from the sound source, but also his own body movements and ongoing position.

Some classic papers on the role of head movement in the context of other non-acoustic cues in sound localization provide a wealth of observations on this topic. (The accompanying chapter by Suzuki et al. (2020) also explores such concerns.) Most notable was early work by Wallach (1940), who observed that head-turning during presentation of a sound stimulus made it possible to distinguish whether a sound arrived from in front or in back of a listener. He noted that when the head was turned to the left, the auditory image associated with a frontal sound source would shift towards the right ear, whereas a dorsal source would shift towards the left. This enables front/rearward distinctions to be made on the basis of head-motion-coupled changes in interaural cues producing variation in the lateral angle of the auditory image. Under conditions in which pinna cues and movement cues indicated incidence from contrasting hemifields, these dynamic interaural cues dominated pinna cues to direction. Wallach also presented such dynamic sound stimuli under conditions in which an illusion of self-rotation was induced by placing stationary subjects inside a revolving screen that filled the visual field. Since their heads were not actually rotating, vestibular cues were absent, and yet listeners experienced self-motion due to these visual cues, and experienced front-to-back reversal when the lateral angle of a frontal sound stimulus was made to shift with head movement as it would were it arriving from the rear.

In another relatively early study, Thurlow and Runge (1967) also investigated the influence of head-rotation on directional hearing, again manually inducing head movements rather than allowing the listener to perform them actively. They examined errors in both azimuth and elevation judgments for a number of types of angular head movement. Without belaboring specifics of the experiments, general results can be summarized as follows: Relative to a condition in which no head movement was allowed, rotation of the head reduced errors in azimuth judgment as expected. However, head-rotation did not significantly reduce errors in elevation judgments. If, alternatively, a subject's head was rolled from side to side while listening (which, in the terminology of the original paper, was called 'pivoted,' as tabulated by Table 4), elevation errors were reduced and azimuth errors were not. This makes sense when considering what happens to the lateral angle of an elevated stationary source when first one ear is dropped closer to the ipsilateral shoulder, and then the other is dropped towards its adjacent shoulder: the lateral shift is the opposite of what is experienced

Table 4 Angular motions of the head (“cocking”)

Euler rotation	Plane	Active semicircular canal	Informal designation	Gesture	Expression
Pitch	Median	Superior, anterior	Tip	Nod	Affirmation, concurrence: “yes”
Yaw	Horizontal	Horizontal, lateral	Rotate	Turn, shake	Denial, contradiction: “no”
Roll	Frontal	Posterior	Pivot	Roll, rock, wag, tilt	Uncertainty, questioning: “maybe”

for stationary sources located well below ear level. When the head was tipped forward and back (facing down then up), neither error rate was reduced significantly, as might be expected from the above analysis, since no lateral shifts would occur.

A more recent study of the relative influence of tipping and pivoting considered perceptual attributes associated with many simultaneous sources, rather than the single source studied in (Thurlow et al. 1967). In a study of immersive spatial impression by Martens and Han (2018), multichannel program material—presented via a 10-channel array of loudspeakers distributed about a hemispherical array that included ‘height channels’—produced a sense of auditory spatial diffuseness comparable to more truly diffuse stimuli presented using twice as many loudspeakers. In contrast, the spatial impression was noticeably less diffuse when the same 10-channel program was reproduced via a more conventional “without-height” loudspeaker array (i.e., employing loudspeakers located only on a single plane near the listener’s ear level). However, this with- versus without-height discrimination in auditory spatial diffuseness was possible in only one of the three head-movement conditions that were tested, and that was the condition in which head-rolling was active.

Considering the geometry involved, it should be clear that above-below disambiguation is enabled by head-rolling-coupled lateral shifts of auditory images along the interaural axis, as demonstrated by Martens et al. (2011). Head-pitching cannot produce analogous disambiguating changes in lateralization for sources that are stable from the world-centric standpoint. For example, if sources are stabilized to remain within the median or even an offset sagittal plane, no lateral shifts occur with head-pitching, but only variation in the HRTF (or ATF) occurs at each ear. Studies have also investigated whether vestibular sensations are strictly required for head-rotation to disambiguate source incidence angles (whether head-turning or -rolling). For example, Lackner (1977) found that illusory self-rotation could be induced by a rotating sound field. He rotated six loudspeakers mounted on a circular frame around the heads of subjects in the dark. Not only did the subjects report that they themselves were rotating and that the sound field was stationary, but they also exhibited compensatory nystagmoid eye movements like those that would occur if they were actually being rotated. More recent studies have examined the compression of auditory space

during rapid head-turns (Leung et al. 2008), confirming that self-motion can have strong effect on auditory scene analysis (Kondo et al. 2012).

3.3 Implications: Multisensory Interfaces

Results of these classic experiments indicate bidirectional interaction between perception of head and body orientation and auditory spatial perception. Such characteristics can be exploited by modern communication systems. For example, besides smartphone-embedded IMUs (inertial measurement units), mobile devices feature various techniques for position sensing. SLAM (simultaneous localization and mapping) techniques—including depth perception, motion tracking, markerless feature tracking, depth from stereo, structure from motion, and area learning—are used in visual position sensing/systems (VPS). Head- and eye-tracking can refine positional awareness. Rich models of both internal and external spaces inform rendering of multichannel, multimodal displays that leverage “sensor fusion” among various sensory modalities. These observations are elaborated in the conclusion to this chapter, which follows the survey of idiomatic soundscape conventions presented in the next section.

4 Procedural Superposition (Logical and Cognitive Conventions): Signals

Having reviewed in previous sections combinations of spatial soundscapes regarding physical (*sound*) and perceptual (*sensation*) considerations, we finally consider procedural models of *signals* that inform soundscape composition and cognitive apprehension, higher-level metaphorical associations with which listeners decode sound fields (Cohen and Martens 2020).

When interacting with virtual displays, explicit mental models aid in the conscious reinterpretation of perceptual impressions. In graphics, non-photorealistic rendering (NPR) describes deliberately expressive distortion or remapping of imagery, for the purposes of art or information visualization, subsuming realism to some superseding goal, such as visual interest or perspicuity, ease in appreciation or understanding. Analogously, auditory displays also admit such relaxation of literal, “sonorealistic” renderings. Shared assumptions, social conventions, and learned idioms compress communication expression. The following subsections describe some “nonsorealistic renderings” (NSR) used to enhance or enable **design**ation, in the semiotic sense of consensual understanding (Jekosch 2005; Sodnik and Tomažič 2015).

4.1 *Separation of Visual and Auditory Perspectives*

Normally, personal audition and vision are thought of as concentric, the respective sensory organs embodied together as they are in one's head. For simple example, movies, video games, and TV shows present audiovisual scenes that resemble what one might plausibly see and hear if one were at the position of the camera and its assumedly coincident microphone. Such conventions extend to spatial media, as cameras might be binocular, visual displays stereographic, microphones stereophonic, and auditory displays binaural. However, telesensory instrumentation allows and encourages independence of modalities.

For architectural walk- (or fly-)through and auralizations (Kleiner et al. 1993), visual and auditory perspectives should match, as if cameras and microphones were integrally deployed. For a concert, an auditory display might be presented "with perspective" (i.e., aligned with visual display), either directly (acquired via coincident microphone) or coherently simulated. However, performed electroacoustic music can be captured by a variety of overhead, on-stage, and spot (accent) microphones, mixed and distributed for monitoring (realtime self-audition by the musicians), sound reinforcement (for live audience), and recording or transmission (for archive or distribution). Mediated concert experiences such as music videos separate visual and auditory perspectives, not insisting that capture, rendering, or simulation of aural perspective match optical position.

Cinematic and gaming idioms also relax literal associations, freely exercising liberty to set aside assumptions of alignment of auditory and visual perspectives. For example, background "score" music (BGM) is non-diegetic (conceptually outside a story space, like narration) and accommodated by such independence. One additionally attends multiple spaces at once, apprehending not only a narrative scene, but also, implicitly, its musical accompaniment. Displacement can reflect temporal offset as well as spatial. For instance, in sort of the same way that a panning camera leads a moving character by framing comfortably ahead, sound of a subsequent scene is often introduced before corresponding visuals.

A viewing audience's or gamer's perspective is privileged, enjoying not only extraordinary optical perspective (cinematography, montage, etc.), but also artificial auditory access, with flexible correspondence among display modalities. The "2nd-person perspective" popular in role-playing games (RPGs) is characterized by such displacement, as the auditory perspective, through which one listens and speaks, is that of an associated avatar, not that of its tethered viewpoint. That is, the human gamer is projected into a puppet or "vactor" (virtual actor), typically viewed from slightly behind and above, through a loosely attached virtual camera. Likewise, projected location of sound associated with such an avatar (generated by a game engine or voice-chat captured from the human pilot) is that of the avatar, not the lagging virtual camera.

4.2 *Separation of Orientation and Location—Directionalization Versus Localization*

Table 5 juxtaposes location, orientation, and position as well as static posture versus dynamic gesture. Space, at least at sensible levels of apprehension, is 3-dimensional, and location is most simply represented numerically by Cartesian triplets (x, y, z) . For example, CAD models are usually represented as vertices, edges, faces, and solids. Such subject-independence is allocentric (Roginska and Geluso 2018) or exocentric, independent of listeners or observers.

For subjective displays, parameterized explicitly or implicitly by standpoint and egocentric direction, polar or spherical coordinates are more convenient than rectilinear coordinates since they are non-homogeneous, in that range (ρ) is dimensionally different from azimuth (θ) and elevation (ϕ). That is, representation or projection of distance is different from horizontal and vertical direction, and can be decoupled.

For object-based encodings, monaural audio streams can be localized for binaural display with ITD, IID, and HRTF-based filtering. Sound objects are most simply directionalized by intensity panning to loudspeakers near a phantom source, but such amplitude- or gain-based techniques cannot realistically convey spatial effects such as early reflections (echoes), modal resonances (standing waves), and late reverberation.

Ordinary surround sound and 5.1 configurations, using channel-based encodings such as those deployed in home theater arrangements, do not usually exploit elevational cues, such as those deliverable via height or overhead (“voice of god”) channels. However, preconditioning signals with ATFs before display through loudspeakers can simulate height cues (Jo et al. 2010; Tanno et al. 2014).

For scene-based encodings such as Ambisonics, each loudspeaker receives its own weighted sum of all channels, spatially sampling spherical harmonic coefficients. An Ambisonic microphone array captures a sound field and encodes a multichannel signal for flexible re-directionalization.

Of the three affine transformations (scaling, rotation, and translation), Ambisonics accommodates only rotation, so such soundfield recordings can be thought of as “prebaked,” forgoing “remixing” flexibility (such as standpoint excursion or interaural baseline adjustability, which scales anatomical signals such as ITD and IID and changes binaural disparity) for optimized rendering.

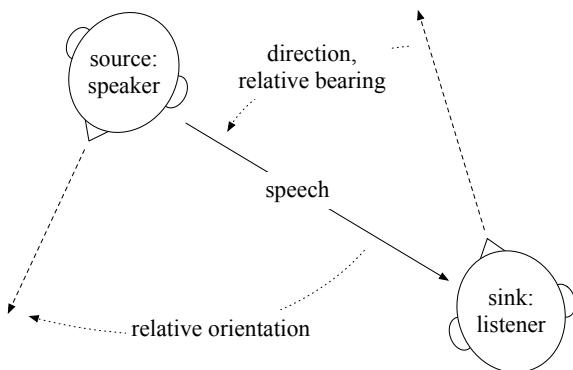
4.3 *Directionality Processing*

Head motion, such as was discussed in the last subsection, is not only like “antenna pointing,” but also “body language,” a kind of display. Situational context, voice intonation, facial expression, gaze and gesture all inform exquisite decoding of proxemic cues. Head gestures as shown earlier in Table 4 are just 1st-order conventions; such communication is rich and subtle. Eye-gaze, which can be approximated from head

Table 5 Physically spatial dimensions: taxonomy of positional degrees of freedom, including cinematographic gestures (assuming right-handed coordinate system, with xy horizontal plane and z gravitational up)

Position (Pose)		Dynamic (Gesture)		
Static (Posture)		Translation (Perturbation)		Along Perpendicular to Plane
Location (Displacement) (Extent)	Scalar	Camera Motion	Directionality (Force)	Axis
lateral, horizontal (breadth, width)	abscissa x	sway track ("crab")	left ↔ right	x
frontal, longitudinal (depth, length)	ordinate y	surge dolly	back (aft): retreat (drag) ↙ front, forth (fore): advance (thrust)	y
vertical (height)	altitude z	heave boom (crane, "ped")	up: ascend (lift) ↓ down: descend (weight)	z
Orientation, Direction, Attitude		About Axis		
Rotation (Spin)		In Plane		
elevation, inclination	ϕ	pitch (tumble, flip) tilt	climb/dive	x
("barrel roll")	ψ	roll (tilt, bank, flop, "Dutch")	left/right	y
azimuth	θ	yaw (whirl, twist) pan	CCW/CW	z
Location and Orientation		Revolution & Rotation		
focal pivot	x, y, θ	phase-locked orbit "spin-around" or inspection		(z)
			(CCW/CW)	(horizontal)

Fig. 6 Direction and orientation: psychoacoustic cues as proxemic social signals. (By “direction” we mean here the relative bearing of a source with respect to a sink, independent of its egocentric rotation; by “orientation” we mean the direction a source is facing.)



orientation, is used for social signaling and can trigger computer-mediated events. Individually apprehended spatial sound tells the eyes where to look, but “gaze indirection” (understanding where someone else is looking), awareness of directed or projected visual attention, alerts conversants about objects of regard. Mouth-emitted sounds are anisotropic, and speech is directional.

As illustrated by Fig. 6, a listener estimates not only direction but also orientation of a talker. Using hints such as ratio of direct-to-indirect intensity and darkening (via low-pass filtering) of utterances, listeners recognize which way a talker is facing, inferring targets of directed address. Symmetrically, talkers are aware of orientation of listeners, and modulate their voices according to appreciation of the listening difficulty of those facing away from them (akin to the Lombard effect, in which talkers strengthen vocalizations in the presence of ambient noise). An aware renderer such as a roomware auditory display is parameterized not only by direction but also orientation of sources relative to sinks, modulating delivered audio streams to convey such fine cues.¹

Sink and source directivity can be modeled by emulating idealizations of microphone receptivity patterns, combinations of omnidirectional (unipolar) and directional (dipolar) radiation as well as sensitivity (Hugonnet and Walder 1998). For typical instance, the Google VR Audio (<https://developers.google.com/vr/ios/spatial-audio>) and Resonance Audio (<https://resonance-audio.github.io/resonance-audio/>) Unity plug-ins model directionality by “alpha” ($0 \leq \alpha \leq 1$) and “sharpness” ($1 \leq \text{sharpness}$). Normalized gain fields are calculated as $|(1 - \alpha) + \alpha \cos(\theta)|^{\text{sharpness}}$, where θ is the relative direction of (for projection or emission) a sink with respect to a source or (for reception or sensitivity) a source w.r.t. a sink, bilinear weighting coefficient α scales directionality, dipole power sharpness exaggerates such non-isotropy, and the absolute value function rectifies

¹A “sink” is the dual of a source, used instead of “listener” to distinguish it from an actual human, including allowing designation of multiple sinks for a single user, as explained in Sect. 4.8 below.

polarity inversion.² When α is zero, the pattern is isotropic (and the sharpness is irrelevant); as α approaches unity, directivity becomes increasingly lobed. “Earshot,” combined radiation and reception, is the product of these for each source \rightarrow sink combination.

Such sensitivity directivity patterns are analogous to clipping frusta of computer graphics rendering. Such hyper-acuity of apprehension or heightened directionality of projection are best suited for AR applications embedded in real world contexts, since purely virtual exposure and receptivity are not constrained by such coarse models as lobed directivity. These are generalized by narrowcasting, described below in Sect. 4.7.

4.4 *Nonrealistic Range-Based Attenuation*

Just as with computer graphics, it is common to introduce both approximate and more complicated models for sound propagation (diffusion, reflection, reverberation, refraction, and diffraction in the presence of obstacles or occluders, dispersion, absorption and scattering) to realize both improved performance and expressive control. Intensity of a point source spherically radiating sound waves naturally observes an inverse square relation with distance, so amplitude gain, a root power quantity proportional to RMS pressure and the square root of intensity, observes a reciprocal (inverse-proportional) relation with range. Distance modulation and estimation of virtual sound sources becomes even sharper if volume control is driven by models that roll-off more rapidly than this physical gain $\propto 1/\rho$ law, where ρ is the distance between source and sink. In contrast, it is sometimes assumed that, in small spaces, amplitude of a reverberant signal changes little with range, and that in large spaces it is roughly proportional to $1/\sqrt{\rho}$ (Pulkki et al. 2011).

Excepting extreme circumstances in spatial sound teleconferencing, such as when a virtual source approaches antipodal position, geotagged sources can be rendered basically horizontally, but with elevation: ignoring spherical curvature of the earth, but allowing relative altitude effects such as mountains and valleys. For many applications, such as conferencing and navigation, it is convenient to separate direction and range, rendering the former faithfully but the later metaphorically or not at all.

For example, realistic display would attenuate most sources below audibility. In everyday experience, even very loud sources are rarely heard beyond a few kilometers, and conversational intensities are normally inaudible beyond tens of meters. With the usual $-6\text{dB}/\text{range doubling}$ attenuation, the level of a typical conversational human speaker, measuring, say, 60 dB SPL at 1 m, weakens a millionfold at 1 km to 0 dB, a nominal auditory threshold, and practical inaudi-

²Similar plug-ins are also offered by other companies, including Facebook (<https://facebookincubator.github.io/facebook-360-spatial-workstation/>), Microsoft (<https://docs.microsoft.com/en-us/azure/cognitive-services/acoustics/what-is-acoustics>), and Yamaha (<https://research.yamaha.com/ja/technologies/vireal/>).

bility occurs even closer because of background noise. Fortunately, utilities for way-finding (such as Microsoft Soundscape (<https://www.microsoft.com/en-us/research/product/soundscape/>)), direction-giving, and conferencing do not need to render sonorealistic range cues.

Besides intensity-controlled loudness, other cues to simulate or suggest distance can be separately modulated (Jot 1999), including initial time-delay gap, the interval between a direct sound and its first reflection; the previously mentioned direct:indirect ratio of the power of direct sound to that of reverberation; motion parallax, subjective shift of a source when the head is moved; and high-frequency attenuation. Nature, including air, is a low-pass filter, and receding sources naturally manifest darkening, thinning of higher frequency components. Direction is usually more important than distance expression, but a fully featured display should allow localization into one's "whisper space" (Villegas and Cohen 2010) to convey such near-field intimacy, such as that evoked by autonomous sensory meridian response (ASMR) programs.

Relatedly, a rendering engine might perform "spotlight mixing," exaggerating loudness of frontal objects assumed to be foci of attention, analogous to foveal rendering in computer graphics. Alternatively, as frontal objects could be assumed to be visible and therefore already conspicuous, rearward objects might be particularly amplified (Bailey 2007), or their auditory position or timbre animated to "catch one's ear." Such "gaze mixing" (<https://docs.microsoft.com/en-us/windows/mixed-reality/spatial-sound>) is a sensory substitution kind of multimodal coordination, which also includes "audio haptics," reactive sounds for touchless interactions, compensating for a lack of force-feedback in virtual displays.

4.5 Extreme Dynamic Range Compression: Location-Indifferent Intensity

Dynamic range is the ratio of the intensities of the strongest and weakest parts of a signal, and range in the sense of source \rightarrow sink distance can be used to attenuate level, distance fall-off. In the limit, compression of dynamic range associated with distance-dependent attenuation approaches range-insensitivity. Separation of orientation and location, including distance independence, allows directionalization without localization. In spatial user interfaces, compass bearings such as "North" are obviously purely directional (like computer graphics directional lights, as opposed to area-, point-, or spot-lights), but even grounded objects with specific locations (such as one's home or office) or characters (such as icons or avatars representing conversants) can project as range-indifferent sources, by normalizing or compressing range-dependent intensities. Sound spatialization can preserve direction but collapse distance.

Affordable systems for immersive photospherical or volumetric visual and stereophonic auditory display represent a popularization of VR-style interfaces. Google Cardboard (<https://arvr.google.com/cardboard/>), the Merge Headset (<https://merge>

[edu.com/headset](https://www.oculus.com/headset)), Oculus Quest (<https://www.oculus.com/quest>), and Samsung Gear VR (<https://www.samsung.com/global/galaxy/gear-vr/>) use sensors for head-tracked binocular display of stereoscopic contents and stereophonic display of spatial audio. Orientation can be tracked by a micro-electro-mechanical system (MEMS) IMU—including gyroscope, accelerometer, and magnetometer—estimating bearing via aggregating sensor fusion, but if location is not tracked (as via GPS or optical tracking), user virtual standpoint is not directly adjusted.

Some scene-based interfaces ignore location and use only orientation. Spatial sound sources can be directionalized without range-parameterized gain modulation. With head-tracking, a subjective soundscape can be counter-rotated, panned to stabilize a scene, but not perturbed. Orientation sensitivity supports location-based sound fields. For example, fields captured or encoded into Ambisonics B-format (with 4 channels) are easily rendered at runtime, down-mixed to a panned stereo pair heard through head-tracked headphones or up-mixed to a real or virtual speaker array.

4.6 Layered Listening and Audio Windowing

Procedural mixing allows user interfaces to algorithmically combine and distribute audio signals. Networked and object-based articulated sources invite audio-level (as opposed to acoustic-level) modulation, and logical layers are a natural model for such composition. Cinema and electronic gaming encourage richly textured soundscapes, including music, sound effects (SFX), narration and dialog channels. Room effects such as echo and reverb can be added by ambience processors.

Graphical compositing, à la Photoshop-style layers, allows various blending modes, articulated effects applied at each phase of the “bit bucket brigade,” a chain of filters like a sequence of guitar effects pedals or a composition of digital effects to enrich expression. Such a cascade is equivalent to a tree of metamixers (Cohen 2015), a dataflow arrangement in which compositing operations are modeled as routing matrix switches with effects applied at each crosspoint—“programmable shaders” fanning-out into amplifiers for a combination of personal and public transducers, headphones and loudspeakers. Multichannel Audio Digital Interface (MADI) (<http://www.aes.org/publications/standards/search.cfm?docID=17>) and Dante (<https://www.audinate.com>) are popular standards for multichannel audio networking and interfaces. Audio middleware and engines such as CSound (<https://csound.com/>), Faust (<http://faust.grame.fr/>), FMOD (<https://fmod.com>), JUCE (<https://juce.com>), Max/MSP (<https://cycling74.com/products/max/>), Pure Data (<http://puredata.info>), Reaktor (<https://www.native-instruments.com/en/products/komplete/synths/reaktor-6/>), SuperCollider (<https://supercollider.github.io>), and Wwise (<https://www.audio-kinetic.com/products/wwise/>) can render auditory scenes.

Audio windowing (Cohen 2016), in analogy to graphically windowing user interfaces (and not to be confused with signal-processing data sequence extraction), treats soundscapes as articulated elements in a composite display (Begault 1994). Spatial soundscapes, like layers in graphical applications or tracks in musical compositions,

can be combined simply by summing, although some scaling (amplification or attenuation), equalization, and other conditioning yields better results. For instance, interior soundscapes might be reverberated, to distinguish them from outdoor scenes. To make a composited soundscape manageable, some sources might be muted or muzzled and some sinks might be deafened or muffled.

As was illustrated by Fig. 4, mixed reality can not only add information to naturally captured scenes, but can also remove information. Interpretation of “XR” can include “Diminished Reality.” Diminished audio reality can be thought of as hiding or masking otherwise apparent auditory scene components, such as engine sounds (as in ANC), objectionable ambient “room tone,” or an unwelcome voice (such as that of a boring interloper). Such “unmixing” suppression of particular sources is the opposite of the “cocktail party effect” (Middlebrooks et al. 2017), whereby particular objects are “heard out” of a cacophonous mix. They are generalized together by auditory source separation, auditory scene analysis (Bregman 1990), and blind source separation.

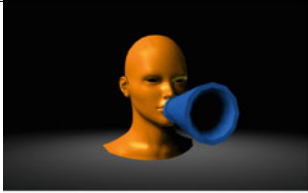





4.7 Narrowcasting: Privacy and Attention Management

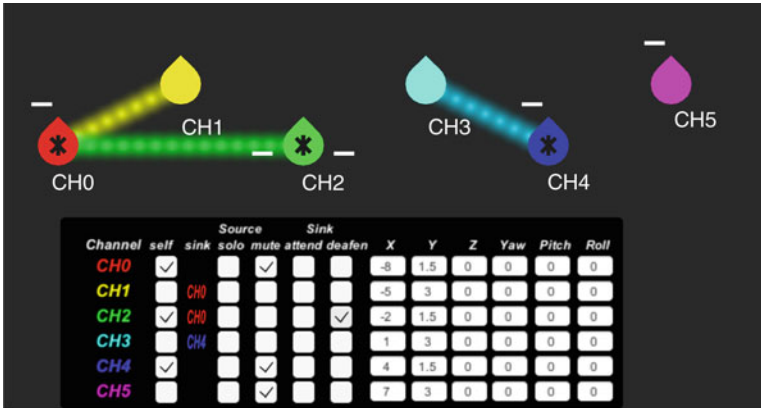
“Privacy” has two interpretations. The first association is that of avoiding “leaks” of confidential information, protecting secrets. The second association is “freedom from disturbance,” not being bothered by interruption. Narrowcasting operations manage privacy in both senses, filtering duplex information through an articulated communication model. In analogy to any-, broad-, multi-, and unicasting, narrowcasting is an idiom for limiting and focusing media streams. Sources and sinks are symmetric duals in virtual spaces. A human user might be represented by both a source and a sink in a groupware environment, or perhaps by multiple instances of such delegates, and both one’s own and others’ sources and sinks can be adjusted for privacy. Sound sources can be explicitly “turned off” by being muted, or implicitly ignored by selecting some others. Similarly, audibility of a soundscape is controlled by embedded sinks, which can be explicitly deafened or implicitly desensitized if other sinks are “attended” (Cohen 2000).

Formalized by the permission scheme expressions shown in Fig. 8, narrowcasting (Alam et al. 2009; Cohen et al. 2009) exposure and distributes attention. Advanced floor control symbology—for chat-spaces, concerts, and conferences—is outlined by Table 6. Modulation of source exposure or sink attention needn’t be “all or nothing”—nimbus (projection) and focus (receptivity) can be respectively partially softened with muzzling and muffling (Cohen 1993)—see Fig. 7.

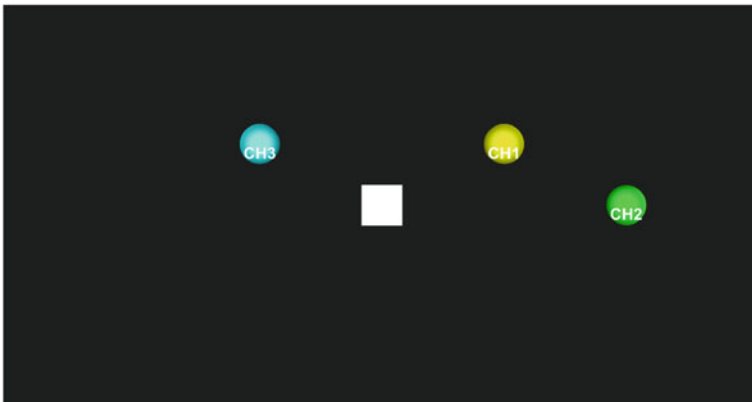
That is, nuanced operations can soften state transition, allowing non-binary control—not just *on-off* but intermediate gains as well—and also signal-processing filter cascades at each opportunity. Narrowcasting attributes can be integrated with spatialization and used for “polite calling” or “awareware,” reflecting sensitivity to one’s availability, like the “*online-offline*” switch of a conferencing service.

Table 6 Narrowcasting for ${}^s\text{OU}_{\text{put}}^{\text{rcc}}$ and ${}^s\text{IN}_{\text{put}}^{\text{k}}$. (Figurative avatars by Julián Villegas.)

		Source (♂)	Sink (♀)
Function		Radiation (effector), emission	Reception (sensor), collection
Level Adjustment		Amplification, Attenuation	Sensitization, Desensitization
Media Direction		OUTPUT (display), production, push	INPUT (control), consumption, pull
Perspective		Object	Subject
Presence Locus		Nimbus (projection, exposure)	Focus (receptivity, sensitivity)
Auditory	Instance	Speaker	Listener
	Transducer	Loudspeaker	Microphone array or dummy-head
	Organs	Mouth	Ears
Enable (spurn)	Include	solo, select	attend (harken)
	Metaphorical Device	Megaphone, loud-hailer, bull-horn	Ear trumpets
	Icon	+ ○	+○+
	Avatar		
Disable (spurn)	Exclude	mute	deafen
	Suppress	Muzzle	Muffle
	Icon	- ○	-○-
	Avatar, own		
	<i>reflexive</i>	(Thumb up)	(Thumbs down)
	Avatar, other		
<i>transitive</i>	(Thumb down)	(Thumbs up)	



(a) Exocentric soundscape interface and “mixels” panel: Three self-identified sinks (tagged with asterisks) audition each other and three other sources. In this somewhat unnatural arrangement, all the sources and sinks face the same direction (upwards in the map). Three of the sources (CH0, CH4, & CH5) are muted (as indicated by frontal minus signs), and one of the sinks (CH2) is deafened (as indicated by laterally straddling minus signs).



(b) “Flattened” soundscape, collapsed around a single listener perspective: Active sources and deafened sinks are partitioned across active sinks, which necessarily coalesce into a singular perspective. Since the autofocus algorithm assigns sources CH1 & CH2 to CH0 and CH3 to CH4, as indicated in the mixels panel above, the soundscape reduces to this subjective arrangement around the notional human listener, iconified by the central white square.

Fig. 7 Dynamic map featuring display and control of spatial sound sources and sinks, including narrowcasting, multipresence, and autofocus (Cohen and Kojima 2018), with contributions by Akane Takeshige, Peter Larson, and Koki Tsuda with Rintarō Satō

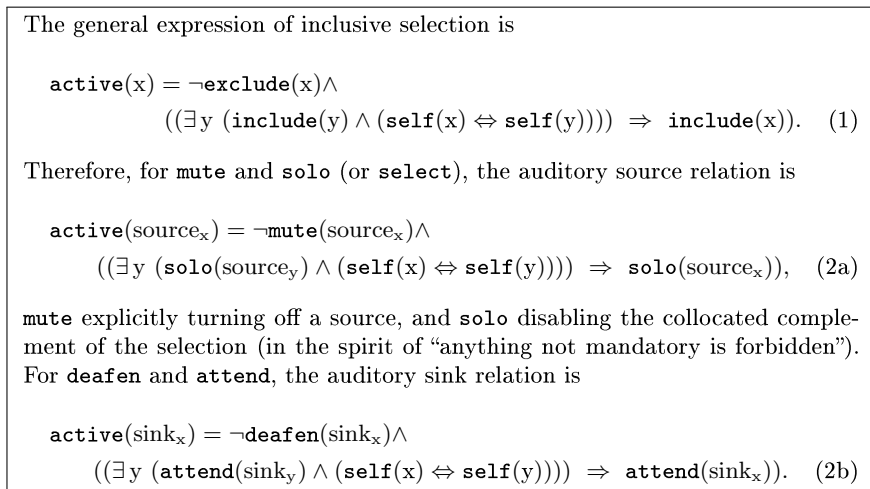


Fig. 8 Formalization of narrowcasting functions in predicate calculus notation, where ‘ \neg ’ means “not,” ‘ \wedge ’ means conjunction (logical “and”), ‘ \exists ’ means “there exists,” ‘ \Rightarrow ’ means “implies,” and ‘ \Leftrightarrow ’ means “is equal to” (mutual implication, “if and only if”). Duality between source and sink operations is strong, and the semantics are analogous: an auditory object is inclusively enabled by default unless, (i) it is explicitly excluded with `mute` (for sources) or `deafened` (for sinks), or, (ii) peers in the same `self`/`non-self` class are explicitly included with `solo`/`select` (for sources) or `attend` (for sinks) when the considered object is not

4.8 Multipresence and “Anyware”

Ordinary correspondence between inhabited bodily apprehension and consciousness is one-to-one, but teleexistence (Tachi 2015) can soften such rigidly focused subjectivity, relaxing the singularity of human experience. Multitasking users want to have presence in several locations at once. For instance, a telephone exemplifies auditory telepresence, projecting conversants to other places besides their corporeal “meat-space” base.

Enriched user interfaces, especially with position-tracking systems or real-time locating systems, encourage multipresence, the inhabiting by representative sources and sinks of multiple locations simultaneously, allowing a human user to designate *doppelgänger* delegates in distributed domains. Exocentric interfaces supporting “out-of-body” experience enable parallel spaces, across which can be designated multiple instances of self-identified avatars (Cohen 1998; Ranaweera et al. 2015) as shown in Fig. 7. “Anyware” multipresence models separate but combinable scenes, allowing users to enjoy selectively distributed attendance.

Direct superposability of soundscapes makes audition especially open to multipresence—unlike vision, which cannot naturally overlay separate scenes. The apparent paradoxes of auditory multipresence can be resolved by an “autofocus” technique that uses Helmholtz reciprocity (exchangeability of sources and sinks) and simulated

precedence effect (perceptual fusion) to disambiguate soundscapes (Cohen and Fernando 2009), like a “snap-to grid.” A soundscape interpreter can resolve source → sink correspondences, directionalizing, localizing, or spatializing each source to its best sink, a function of respective and mutual direction and orientation, directionality, and range.

4.9 Implications: Nonsonorealistic Rendering and Multimodal Cognition

Exploiting multimodal sensation and mental models of situations and environments, convention and idiom can tighten apprehension of a scene, using metaphor and relaxed expectation of sonorealism to enrich communication. Communication culture is not innate but learned. Listening is not a one-off event, but continuous experience. Sound displays use acquired associations, rather than direct emulation of natural phenomena. An assumed sophistication of listeners decoding nonliteral displays admits an acceptance of plausible but nonveridical cues.

Many situations do not call for an auralization-style re-creation of a particular soundscape but instead are best served by some kind of metaphorical space. Practical auditory conventions such as those described by this section refine expression. For instance, by using an audio windowing system as a mixing console, a multidimensional pan-pot, users and applications determine rich parameters to compile source and sink positions and their environments, rendering as a distributed diffuser or spatial sound stager. Presence is more important than fidelity, audiophilic predilection for “absolute sound” or perceived need for Master Quality Authenticated (MQA; <http://mqa.co.uk>) streaming notwithstanding.

Purely auditory displays hardly exist. Normal physical environments ensure that ordinary events are perceived multimodally. Spatial sound cues are aspects of a rich ecology of environment-embedded signs. Almost always, “in the wild,” visual cues and other context complement projected auditory scenes. Soundscapes are not apprehended “in a vacuum”: some map, conventional understanding, or at least situation awareness aids decoding. Multimodal interfaces empower overlapping displays.

Cognitive processes can resolve otherwise confusing soundscapes. For instance, a flashing light (as on an active smart speaker, or the “Lyric Speaker,” (<https://lyric-speaker.com>) which animates words in *karaoke*-style sync with songs) can disambiguate conflicting cues. Listeners are inclined to be forgiving, suspending not only disbelief but also insistence on sonorealism, so sonic situations can be efficiently communicated. Mental models are used to interpret multimodal events, including those generated by non-literal displays. For instance, independence of location and orientation can flatter and “flatten” multipresent auditory localization. An advantage of separating translation and rotation is that directionalizability can be preserved even across multiple frames of reference. Such distributed presence can be coupled with vehicle or position tracking. Moving can twist (but deliberately not shift) multiple

representations, maintaining consistent proprioceptive alignment of overlaid sound sources.

5 Crowds and Clouds: Final Thoughts and Conclusions

5.1 *Ubicomp and IoT: Extreme Sound Reinforcement*

Ordinary rooms often host electronic appliances such as TVs, desktop and laptop computers, game consoles and controllers, smart speakers, as well as tablets, and smartphones of “second screening” (multitasking) occupants, who might also have HMDs or smart glasses for XR, wearable computers (such as smart watches and hearables), PSAPs and hearing aids. These multitudes of speakers and microphones, displays and sensors, can be integrated by roomware.

In ubicomp environments, generally multiple users must be accommodated. Urban computing offers even broader challenges and opportunities: public signage and auditory displays can serve AR messages to tracked users. A distributed ecosystem of electronic devices defies top-down management but invites bottom-up coordination. Privacy, attention, and sensitivity parameterize rendering of soundscapes. Delegated by human users, software agents and intelligent assistants will negotiate private and collective access to resources. Transducers of AI-infused networked appliances can work in concert with personal “awareable” devices to optimize personal and public experience. Syndicates of groupware interfaces will pool crowd-sourced data and share displays: mediated social sensing and signaling.

In an “ABC” (always best connected) world, persistent chat-spaces are expected: selectively continual connectivity with one’s family, friends, and colleagues. Aware interfaces infer user receptivity, tuning an environment by automatically adjusting displays of all types to reward attention. Activity sensors, position trackers, and monitors cooperate to optimize comfort, efficiency, and productivity. IoT-style smart speakers should be situationally aware, using amalgamated sensing—microphones, cameras (including thermal and infrared sensors), mo-cap, EEG, and fitness trackers and biosensors (capturing microexpressions of voice, gaze, body language, pupil dilation, heartbeat and pulse variability, galvanic skin response, body heat, etc.)—to gauge mood, empathetically adjusting soundscapes to support users (Crum 2019).

Compiling a heterogeneous display, for listeners in arbitrary positions, across speakers of various sizes, orientations, directivities, spectra, acoustic intensities, and irregular and dynamic arrangement is endlessly challenging: extreme sound reinforcement. However, opportunistic networked managers (Choi et al. 2016) can exploit disparate devices for enriched presentation, carving out “sound zones.” Reflexive display-and-capture systems can be used to calibrate diffusion in a “closed loop,” like that used by structured light sensing. For instance, roomware might arrange to ‘borrow’ or ‘lease’ nearby sensors and effectors to adjust parameters. Representative contemporary applications demonstrate such cyberphysical cooper-

ation between speakers and microphones and suggest the potential of such symbiosis: “Chirp” (<https://chirp.io>) and “Google Tone” (<https://chrome.google.com/webstore/detail/google-tone/nnckehldicaciogcbchegobnafnjkcne>) distribute URLs to nearby computers audibly (“data-over-sound”); “Ultrasonic Recognition” (<http://www.lankasolution.com/ar365-usr-ultra-sonic-recognition/>) embeds tags in audio tracks; and “AmpMe” (<http://ampme.com>) and “Tune Mob” (<https://itunes.apple.com/developer/tunemob/id680664869>) manage network-synchronized distributed music display. Audio steganography can embed “side-channel” information as subliminal, ultrasonic, or otherwise inaudible acoustic signals.

5.2 AI-Empowered Conversational Agents

Besides mobile telephony, so-called “smart speakers,” which also integrate microphone arrays and often lights or fuller displays, feature internet services for conversational interfaces backed by AI for information or control. Emergent qualities of networked sensors and the high bandwidth and low latency of wireless systems such as that promised by 5G, 5th-generation cellular networks, recall the blending of fixed-mobile convergence (FMC). As the processing is mostly on-line, intelligence cannot be attributed to the loudspeaker itself: the network makes locality of computation seamless or “cloudy.” We extend ourselves with distributed systems, and the network stretches to embrace us cyberspatially.

Such IoT devices represent an interpolation between robots and chatbots, transactional and conversational virtual assistants. Appliances, even with wireless data connections, are usually powered and fixed, but ambulatory electronic pets and consumer robots—including socially assistive models and hospitality-service bots (such as Sony Aibo (<https://us.aibo.com>), Honda Asimo (<http://asimo.honda.com>), SoftBank Pepper (<https://www.softbank.jp/en/robot/>), and Sharp RoBoHon (<https://robohon.com/global/>))—detecting and responding to human emotions, represent self-locomotive loudspeaker platforms with telepresence capability.

Acoustic devices can be wireline or wireless, spanning continua of data- and power-cordlessness: **Fixed**, as by normal loudspeakers; **Tethered**, as by many HMDs; **Bounded**, as with zones for near field communication (NFC) and area networks such wireless local area networks (WLANS) and near-me area networks (NANS), including those of Bluetooth, Wi-Fi and WiGig; and **Free-roaming**, as with cellular coverage.

Voice interfaces feature speech recognition (SR) and text-to-speech (TTS), with increasingly natural sounding synthesis, allow rendering of textual sources as auditory sources, a synaesthetic transcoding. The renaissance of machine learning and AI includes advances in big data and deep learning, for speech interpretation, machine translation, conversational intelligence, and multilingual TTS. “Vocal emotion recognition” can characterize mood from speech, using such microexpressive cues as voice dynamics, tone, timing, and metalinguals. AI can be applied to situation awareness, estimating social conditions such as user sensitivity (distractibility, attention, fatigue, multitasking, “flow”), including support functions such as face, speech and speaker

recognition; optical character recognition (OCR); natural language processing (NLP); and “*kansei* (affective) engineering” sentiment analysis.

Enabled by the confluence of sensing, connectivity, computation, and machine intelligence, user recognition and characterization allow provision of personalized media and listening zones. The “quantified self” domain includes audiometric customization and individualization of ATFs. Public loudspeakers are usually around the periphery of a room—often at the walls, sometimes on the ceiling, rarely in the floor—but smart speakers among and amidst people can complement traditional loudspeakers, and along with personal displays, contribute to integrated mobile-ambient interfaces for immersive experience, taking “theatre-in-the-round” and turning it envelopeingly inside-out. Paralleling FMC, “glocal” interfaces can leverage both personal devices and shared resources. For control, smartphone-sensed orientation and GPS-like tracking can be combined with parameters such as layering and narrow-casting attributes. For display, smartphone and tablet screens can be extended by cooperative roomware lights and screens, and headphones and hearables can be augmented by speaker arrays.

5.3 *Late Binding of Soundscape Staging: Runtime Determination of Synthesis, Filtering, Spatialization, and Multimodal Rendering*

Spatial sound systems handle three different kinds of audio encodings, namely,

Channel-based, associated with fixed (“bed”) display configurations (headphones, stereo speakers, home theater layouts, theatrical arrangements, etc.) including matrix encodings,

Scene-based, such as Ambisonics recordings and streams that capture sound fields at particular locations

Object-based, associating streams with particular objects in a scene (human speakers, musical instruments, acoustic events), and assuming that an audio renderer will directionalize or spatialize these tracks for a parameterized display.

Audio sources for games (Collins 2008) and simulations have historically been associated with prerecorded files, but more richly parameterized applications and social media drive a shift to dynamic media streams, including physical modeling, procedural audio, algorithmic music, voice-chat, and, inevitably and imminently, “deepfake” photo- and sonorealistic multimedia. The parallel trend is away from assumed fixed loudspeaker locations and towards expectation that material will be rendered to whatever is available at the display end of the chain. As attention shifts away from prepared media towards online experiences, the process of mixing changes: instead of aggregation into “stems,” raw audio tracks are pushed into dynamic rendering, configured by metadata object positions and realtime tracking. Rather than baking virtual sources into transducer channels, which is a kind of

rigid compilation, sources are rendered and diffused at runtime, accommodating circumstances and exploiting opportunities. Parameterization by “late binding” display arrangement is a kind of dynamic projection mapping, configuring signal-processing to match particular loudspeaker and headphone resources and configurations.

Such freestyle improvisation lacks the broad consistency of cinematic standards such as Auro-3D (<https://www.auro-3d.com>), DTS:X (<https://dts.com/dtsx>), and Dolby Atmos (<https://www.dolby.com/us/en/brands/dolby-atmos.html>), but is potentially richer and is inherently future-proof. Dolby AC-4 (<https://www.dolby.com/us/en/technologies/AC-4.html>) combines channel- and object-based models, and DTS MDA, Multi Dimensional Audio, is a kind of interpolation between channel- and object-based encoding, with object-based channels mapped to theatrical speakers at installation time. Encoding standards for channel-, scene-, and object-based models were reviewed by Cohen and Villegas (2016). The MPEG-H (<https://www.mpeg-h.com/en/home/>) 3D Audio (<https://mpeg.chiariglione.org/standards/mpeg-h/3d-audio>) and the ITU ADM (Audio Definition Model; https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.2125-0-201901-I!!PDF-E.pdf) standards integrate these models. For typical instance, object-based foreground spatialization can be rendered atop both channel-based stereo (non-diegetic) BGM and scene-based Ambisonic “sweetening” atmospheric background.

Synergies among components arise even for someone alone in a room. Such mutual support includes ducking during voice chats to attenuate backgroundable media; using smartphones and smart speakers to reinforce or articulate cinematic soundtracks and conferencing channels; and using IoT addressability to integrate distributed displays (such as speakers and lights) and sensors (such as microphones and cameras).

Media device orchestration (Francombe et al. 2018) uses ad hoc arrays of appliances to augment apprehension. In the parlance of media presentation, a responsive framework serves dynamic content through an adaptive heterogeneous display. Articulation and comodulation of parameters can coordinate audio and visual displays to accommodate attention, mood, and circumstances. Synchronicity of complementary cross-modal signals—such as moving lips or flashing light, or a map or Gestalt mental model—can disambiguate otherwise indeterminate cues, or even override preliminary interpretation. Confederation of information appliances, sharing data and capabilities, can enhance awareness, expressiveness, and experience.

To recapitulate, conversation, lectures, phone calls, music, television, and announcements inundate us with sonic signals—purely acoustic, electroacoustic. These auditory stimuli comprise overlaid and attentionally oversaturated spatial sound fields, engulfing listeners cacophonously. Sound is mixed acoustically, perceptually, and cognitively—roughly and respectively associated with the air, ear, and brain—corresponding to the three kinds of spatial soundscape superposition described in this chapter, that is, physical transmission (sound), perceptual apprehension (sensation), and procedural interpretation (signal).

Together they span our anticipation for the future of auditory interfaces: heterogeneous, personal and public speakers awarably integrated into multimodal duplex interfaces leveraging idiomatic and metaphorical conventions.

Acknowledgements We thank Yōiti Suzuki for his valuable comments and suggestions. This chapter has been reviewed by two anonymous experts.

References

- Alam, S., M. Cohen, J. Villegas, and A. Ahmed. 2009. Narrowcasting in SIP: Articulated privacy control. In *SIP Handbook: Services, Technologies, and Security of Session Initiation Protocol*, ed. S.A. Ahson, and M. Ilyas, 323–345. Boca Raton: CRC Press, Taylor & Francis. Chap. 14. <https://doi.org/10.1201/9781315218939>.
- Bailey, R. 2007. Spatial emphasis of game audio: How to create theatrically enhanced audio. In *Audio Anecdotes III*, ed. K. Greenebaum, and R. Barzel, 399–406. Wellesley: A K Peters/CRC Press. <https://doi.org/10.1201/9781439864869>.
- Bauk, J.L., and D.H. Cooper. 1996. Generalized transaural stereo and applications. *Journal of the Audio Engineering Society* 44 (9): 683–705. <http://www.aes.org/e-lib/browse.cfm?elib=7888>.
- Begault, D.R. 1994. *3-D Sound for Virtual Reality and Multimedia*. Boston: Academic Press. ISBN 978-0120847358
- Blauert, J. 2012. A perceptionist’s view on psychoacoustics. *Arch. Acoust.* 37 (3): 365–371. <https://doi.org/10.2478/v10168-012-0046-z>.
- Blauert, J. 2017. “Reading the World with Two Ears” Keynote at Int. Congress on Sound and Vibration, London. <https://www.youtube.com/watch?v=p1kDtggmTdw>.
- Blauert, J., D. Kolossa, K. Obermayer, and K. Adiloğlu. 2013. Further challenges and the road ahead. *Modern Acoustics and Signal Processing*, 477–501. Berlin: Springer. https://doi.org/10.1007/978-3-642-37762-4_18. Chap. 18.
- Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge: MIT Press.
- Brimijoin, W.O., and M.A. Akeroyd. 2012. The role of head movements and signal spectrum in an auditory front/back illusion. *i-Perception* 3 (3): 179–182. <https://doi.org/10.1068/i7173sas>.
- Broll, W., I. Lindt, I. Herbst, J. Ohlenburg, A.-K. Braun, and R. Wetzel. 2008. Towards next-gen mobile AR games. *Computer Graphics and Animation* 28 (4): 40–48. <https://doi.org/10.1109/MCG.2008.85>.
- Cabrera, D., H. Sato, W. Martens, and D. Lee. 2009. Binaural measurement and simulation of the room acoustical response from a person’s mouth to their ears. *Acoustics Australia* 37: 98–103.
- Choi, J.-W., B.J. Cho, and I. Shin. 2016. Toward the holographic reconstruction of sound fields using smart sound devices. *IEEE MultiMedia* 23 (3): 64–74. <https://doi.org/10.1109/MMUL.2016.46>.
- Choueiri, E. 2018. Binaural audio through loudspeakers, in Roginska and Geluso. <https://doi.org/10.4324/9781315707525>. Chap. 5.
- Cohen, M. 1993. Throwing, pitching, and catching sound: Audio windowing models and modes. *IJMMIS: Journal of Person-Computer Interaction* 39 (2): 269–304. <https://doi.org/10.1006/imms.1993.1062>.
- Cohen, M. 1998. Quantity of presence: Beyond person, number, and pronouns. In *Cyberworlds*, ed. T.L. Kunii, and A. Luciani, 289–308. Tokyo: Springer. https://doi.org/10.1007/978-4-431-67941-7_19. Chap. 19.
- Cohen, M. 2000. Exclude and include for audio sources and sinks: Analogs of mute & solo are deafen & attend. *Presence: Teleoperators and Virtual Environments* 9 (1): 84–96. <https://doi.org/10.1162/1054746005666637>.
- Cohen, M. 2015. Hierarchical narrowcasting. In *Proceedings of HCII: International Conference on Human-Computer Interaction– DAPI: International Conference on Distributed, Ambient and Pervasive Interactions*, ed. N. Streitz and P. Markopoulos, 274–286. Los Angeles: LNCS 9189. https://doi.org/10.1007/978-3-319-20804-6_25.

- Cohen, M. 2016. Dimensions of spatial sound and interface styles of audio augmented reality: Whereware, wearware, & everywhere. In *Fundamentals of Wearable Computers and Augmented Reality*, ed. W. Barfield, 277–308. Mahwah: CRC Press. <https://doi.org/10.1201/b18703>. Chap. 12.
- Cohen, M., and O.N.N. Fernando. 2009. Awareware: Narrowcasting attributes for selective attention, privacy, and multipresence. In *Awareness Systems: Advances in Theory, Methodology and Design*, ed. P. Markopoulos and W. Mackay, 259–289. London: Springer. <https://doi.org/10.1007/978-1-84882-477-5>. Chap. 11.
- Cohen, M., O.N.N. Fernando, U.C. Dumindawardana, and M. Kawaguchi. 2009. Duplex narrowcasting operations for multipresent groupware avatars on mobile devices. *IJWMC: International Journal of Wireless and Mobile Computing* 3 (4): 280–287. <https://doi.org/10.1504/IJWMC.2009.029348>.
- Cohen, M., and H. Kojima. 2018. Multipresence and autofocus for interpreted narrowcasting. In *AES: Audio Engineering Society International Conference on Spatial Reproduction—Aesthetics and Science*, Tokyo. <http://www.aes.org/e-lib/browse.cfm?elib=19653>.
- Cohen, M., W.L. Martens. 2020. Spatial soundscape superposition, Part II: Signals and systems. *Acoustical Science and Technology* 41.1 (Jan. 2020). ed. by Masato Akagi, Masashi Unoki, and Yoshifumi Chisaki. JASJ 76 (1): 297–307. ISSN: 1347-5177, 1346-3969, 0369-4232. <https://doi.org/10.1250/ast.41.297>.
- Cohen, M., and J. Villegas. 2016. Applications of audio augmented reality: Wearware, everywhere, anywhere, & awareware. In *Fundamentals of Wearable Computers and Augmented Reality*, 2nd ed, ed. W. Barfield, 309–330. Mahwah: CRC Press. <https://www.taylorfrancis.com/books/9780429192395>. Chap. 13.
- Collins, K. (ed.). 2008. *Game Sound*. Cambridge: MIT Press. <https://doi.org/10.7551/mitpress/7909.001.0001>. ISBN 978-0-262-03378-7.
- Crum, P. 2019. Here come the hearables: Technology tucked inside your ears will augment your daily life. *IEEE Spectrum* 56 (5): 38–43. <https://doi.org/10.1109/MSPEC.2019.8701198>.
- Francombe, J., J. Woodcock, R.J. Hughes, R. Mason, A. Franck, C. Pike, T. Brookes, W.J. Davies, P.J.B. Jackson, T.J. Cox, F.M. Fazi, and A. Hilton. 2018. Qualitative evaluation of media device orchestration for immersive spatial audio reproduction. *Journal of the Audio Engineering Society* 66 (6): 414–429. <http://www.aes.org/e-lib/browse.cfm?elib=19581>.
- Hartmann, W.M. 1999. *Signals, Sound, and Sensation*. New York: AIP Press.
- Herder, J., and M. Cohen. 2002. The helical keyboard: Perspectives for spatial auditory displays and visual music. *JNMR: Journal of New Music Research* 31 (3): 269–281. <https://doi.org/10.1076/jnmr.31.3.269.14180>.
- Hugonnet, C., and P. Walder. 1998. *Stereophonic Sound Recording: Theory and Practice*. Chichester: Wiley. ISBN 978-0471974871.
- Inoue, A., Y. Ikeda, K. Yatabe, and Y. Oikawa. 2017. Three-dimensional sound-field visualization system using head mounted display and stereo camera. In *Proceedings of ASA Meetings on Acoustics*, vol. 29. <https://doi.org/10.1121/2.0000381>.
- Jekosch, U. 2005. Assigning meaning to sounds—semiotics in the context of product-sound design. In *Communication Acoustics*, ed. J. Blauert. Berlin: Springer. https://doi.org/10.1007/3-540-27437-5_8. Chap. 8.
- Jo, H., W.L. Martens, Y. Park, and S. Kim. 2010. Confirming the perception of virtual source elevation effects created using 5.1 channel surround sound playback. In *VRCAI: Proceedings of International Conference on Virtual-Reality Continuum and Its Applications in Industry*, 103–110. Seoul: ACM. <https://doi.org/10.1145/1900179.1900200>.
- Jot, J.-M. 1999. Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia Systems* 7 (1): 55–69.
- Kawaura, J., Y. Suzuki, F. Asano, and T. Sone. 1991. Sound localization in headphone reproduction by simulating transfer functions from the sound source to the external ear. *Journal of the Acoustical Society of Japan (E)* 12 (5): 203–216. <https://doi.org/10.1250/ast.12.203>.

- Kendall, G. 2010. Spatial perception and cognition in multichannel audio for electroacoustic music. *Organised Sound* 15 (3): 228–238. <https://doi.org/10.1017/S1355771810000336>.
- Kim, C., R. Mason, and T. Brookes. 2013. Head movements made by listeners in experimental and real-life listening activities. *Journal of Audio Engineering Society* 61 (6): 425–438. <http://www.aes.org/e-lib/browse.cfm?elib=16833>.
- Kleiner, M., B.-I. Dalenbäck, and P. Svensson. 1993. Auralization— an overview. *Journal of Audio Engineering Society* 41 (11): 861–875. <http://www.aes.org/e-lib/browse.cfm?elib=6976>.
- Kondo, H.M., D. Pressnitzer, I. Toshima, and M. Kashino. 2012. Effects of self-motion on auditory scene analysis. *Proceedings of the National Academy of Sciences* 109 (17): 6775–6780.
- Lackner, J.R. 1977. Induction of nystagmus in stationary subjects with a rotating sound field. *Aviation, Space and Environmental Medicine* 48 (2): 129–131.
- Lackner, J.R. 1983. Influence of posture on the spatial localization of sound. *Journal of Audio Engineering Society* 31 (9): 650–661. <http://www.aes.org/e-lib/browse.cfm?elib=18987>.
- Leung, J., D. Alais, and S. Carlile. 2008. Compression of auditory space during rapid head turns. *Proceedings of the National Academy of Sciences* 105 (17): 6492–6497. <https://doi.org/10.1073/pnas.0710837105>.
- Lossius, T., P. Baltazar, and T. de la Hogue. 2009. DBAP—Distance-based amplitude panning. In *Proceedings of the International Computer Music Conference, ICMC*, (Aug. 16–21, 2009) Montréal, Quebec, Canada. <https://hdl.handle.net/2027/spo.bbp2372.2009.111>.
- Lyon, E., ed. 2016. *Computer Music J.: High-Density Loudspeaker Arrays, Part 1: Institutions*, 40, https://doi.org/10.1162/COMJ_e_00388.
- Lyon, E., ed. 2017. *Computer Music J.: High-Density Loudspeaker Arrays, Part 2: Spatial Perception and Creative Practice*, 41, https://doi.org/10.1162/COMJ_a_00403.
- Macpherson, E.A. 2013. Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation. In *Proceedings of Meetings on Acoustics*, Vol. 19, p. 050131. <https://doi.org/10.1121/1.4799913>.
- Martens, W., S. Sakamoto, L. Miranda, and D. Cabrera. 2013. Dominance of head-motion-coupled directional cues over other cues during walking depends upon source spectrum. In *Proceedings of Meetings on Acoustics*, Vol. 19, p. 050129. <https://doi.org/10.1121/1.4800124>.
- Martens, W.L., D. Cabrera, and S. Kim. 2011. The ‘phantom walker’ illusion: Evidence for the dominance of dynamic interaural over spectral directional cues during walking. In *Principles and Applications of Spatial Hearing*, ed. Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and H. Kato, 81–102. Singapore: World Scientific. <https://doi.org/10.1142/7674>.
- Martens, W.L., and M. Cohen. 2020. Spatial soundscape superposition, Part I: Subject motion and scene sensibility. In *Acoustical Science and Technology* 41.1 (Jan. 2020). ed. by Masato Akagi, Masashi Unoki, and Yoshifumi Chisaki. *JASJ* 76 (1): 288–296. ISSN: 1347-5177, 1346-3969, 0369-4232. <https://doi.org/10.1250/ast.41.288>.
- Martens, W.L., Y. Han. 2018. Discrimination of auditory spatial diffuseness facilitated by head rolling while listening to ‘with-height’ versus ‘without-height’ multichannel loudspeaker reproduction. In *Proceedings of Audio Engineering Society International Conference on Spatial Reproduction*, Tokyo. <http://www.aes.org/e-lib/browse.cfm?elib=19608>.
- Marui, A., and W.L. Martens. 2006. Spatial character and quality assessment of selected stereophonic image enhancements for headphone playback of popular music. In *AES: Audio Engineering Society Conv. (120th Conv.)*, Paris. <http://www.aes.org/e-lib/browse.cfm?elib=13622>.
- Middlebrooks, J.C., J.Z. Simon, A.N. Popper, and R.R. Fay (eds.). 2017. *The Auditory System at the Cocktail Party*. Cham: Springer. <https://doi.org/10.1007/978-3-319-51662-2>.
- Milgram, P., and H. Colquhoun Jr. 1999. A taxonomy of real and virtual world display integration. In *Mixed Reality: Merging Real and Virtual Worlds*, ed. Y. Ohta and H. Tamura, 5–30. Omsha: Springer. Chap. 1. ISBN 978-3-642-87514-4
- Ochiai, Y., T. Hoshi, and I. Suzuki. 2017. Holographic whisper: Rendering audible sound spots in three-dimensional space by focusing ultrasonic waves. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, 4314–4325. New York. <https://doi.org/10.1145/3025453.3025989>.

- Pastore, M.T., Y. Zhou, and W.A. Yost. 2020. Cross-modal and cognitive processes in sound localization. In *The Technology of Binaural Understanding*, eds. J. Blauert and J. Braasch, 315–350. Cham, Switzerland: Springer. Chap. 12. https://doi.org/10.1007/978-3-030-00386-9_12.
- Pereira, F., and W.L. Martens. 2018. Psychophysical validation of binaurally processed sound superimposed upon environmental sound via an unobstructed pinna and an open-ear-canal earspeaker. In *Proceedings of Audio Engineering Society International Conference on Spatial Reproduction*, Tokyo. <http://www.aes.org/e-lib/browse.cfm?elib=19626>.
- Perrott, D.R., H. Ambarsoom, and J. Tucker. 1987. Changes in head position as a measure of auditory localization performance: Auditory psychomotor coordination under monaural and binaural listening conditions. *Journal of the Acoustical Society of America* 82 (5): 1637–1645.
- Plenge, G. 1974. On the difference between localization and lateralization. *Journal of the Acoustical Society of America* 56: 944–951. <https://doi.org/10.1121/1.1903353>.
- Pulkki, V. 1997. Virtual source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society* 45 (6): 456–466.
- Pulkki, V., T. Lokki, and D. Rocchesso. 2011. Spatial effects. In *DAFX: Digital Audio Effects*, 2nd ed, ed. U. Zölzer, 139–184. West Sussex: Wiley. <https://doi.org/10.1002/9781119991298.ch5>. Chap. 5.
- Ranaweera, R., M. Cohen, and M. Frishkopf. 2015. Narrowcasting and multipresence for music auditioning and conferencing in social cyberworlds. Presence: Teleoperators and Virtual Environments 24 (3): 220–242, https://doi.org/10.1162/PRES_a_00232.
- Roginska, A., and P. Geluso (eds.). 2018. *Immersive Sound: The Art and Science of Binaural and Multi-channel Audio*. Routledge: Taylor & Francis. <https://doi.org/10.4324/9781315707525>.
- Satongar, D., C. Pike, Y.W. Lam, and A.I. Tew. 2015. The influence of headphones on the localization of external loudspeaker sources. *Journal of the Audio Engineering Society* 63 (10): 3–19. <https://doi.org/10.17743/jaes.2015.0072>.
- Seldess, Z. 2014. “MIAP: Manifold-interface Amplitude Panning in Max/MSP and Pure Data” in *Audio Engineering Society Convention 137*, Los Angeles, <http://www.aes.org/e-lib/browse.cfm?elib=17435>.
- Shaw, E.A.G. 1982. 1979 Rayleigh medal lecture: The elusive connection. In *Localization of Sound: Theory and Applications*, ed. R. Gatehouse, 13–29. Groton: Amphora Press.
- Sodnik, J., and S. Tomažič. 2015. *Spatial Auditory Human-Computer Interfaces*. Cham, Switzerland: Springer. <https://doi.org/10.1007/978-3-319-22111-3>.
- Streicher, R., and F.A. Everest. 2006. *The New Stereo Soundbook*, 3rd ed. Pasadena: Audio Engineering Associates. ISBN 978-0-9665162-1-0.
- Sullenberger, R.M., S. Kaushik, and C.M. Wynn. 2019. Photoacoustic communications: Delivering audible signals via absorption of light by atmospheric H₂O. *Optics Letters* 44 (3): 622–625. <https://doi.org/10.1364/OL.44.000622>.
- Suzuki, Y., A. Honda, Y. Iwaya, M. Ohuchi, and S. Sakamoto. 2020. Binaural display supporting active listening: Perceptual bases and welfare applications initial proposal: Training of spatial perception with binaural displays supporting active listening. In *The Technology of Binaural Understanding*, eds. J. Blauert and J. Braasch, 665–695. Cham, Switzerland: Springer and ASA Press. Chap. 22. https://doi.org/10.1007/978-3-030-00386-9_22.
- Tachi, S. 2015. *Telexistence*, 2nd ed. Singapore: World Scientific Publishing Company.
- Tanno, K., A. Saji, and J. Huang. 2014. A 3D sound generation system with horizontally arranged five-channel loudspeakers. *IEICE Transactions on Information and Systems* 2 (J97-D(5)): 1044–1052.
- Thurlow, W.R., J.W. Mangles, and P.S. Runge. 1967. Head movements during sound localization. *Journal of the Acoustical Society of America* 42 (2): 489–493. <https://doi.org/10.1121/1.1910605>.
- Thurlow, W.R., and P.S. Runge. 1967. Effect of induced head movements on localization of direction of sounds. *Journal of the Acoustical Society of America* 42 (2): 480–488. <https://doi.org/10.1121/1.1910604>.
- Toole, F.E. 1969. In-head localization of acoustic images. *Journal of the Acoustical Society of America* 48: 943–949. <https://doi.org/10.1121/1.1912233>.

- Villegas, J., and M. Cohen. 2010. Hrir: Modulating range in headphone-reproduced spatial audio. In *VRCAL: Proceedings of International Conference on Virtual-Reality Continuum and Its Applications in Industry*, Seoul. <https://doi.org/10.1145/1900179.1900198>.
- Wallach, H. 1940. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology* 27: 339–368. <https://doi.org/10.1037/h0054629>.
- Wang, D. 2017. Deep learning reinvents the hearing aid. *IEEE Spectrum* 54 (3): 32–37. <https://doi.org/10.1109/MSPEC.2017.7864754>.
- Wenzel, E.M., D.R. Begault, and M. Godfroy-Cooper. 2018. Perception of spatial sound. In Roginska and Geluso (2018), 5–39. <https://doi.org/10.4324/9781315707525>. Chap. 1.
- Wersényi, G., and J. Wilson. 2015. Evaluation of head movements in short-term measurements and recordings with human subjects using head-tracking sensors. *Acta Technica Jaurinensis* 8 (3): 218–229.
- Wolfe, J. 2018. From idea to acoustics and back again: the creation and analysis of information in music. *Substantia* 1: 77–91. <https://doi.org/10.13128/Substantia-42>.

Evaluating Aural-Scene Quality and Speech Understanding

Binaural Evaluation of Sound Quality and Quality of Experience



Alexander Raake and Hagen Wierstorf

Abstract The chapter outlines the concepts of *Sound Quality* and *Quality of Experience* (QoE). Building on these, it describes a conceptual model of sound quality perception and experience during active listening in a spatial-audio context. The presented model of sound quality perception considers both bottom-up (signal-driven) as well as top-down (hypothesis-driven) perceptual functional processes. Different studies by the authors and from the literature are discussed in light of their suitability to help develop implementations of the conceptual model. As a key prerequisite, the underlying perceptual ground-truth data required for model training and validation are discussed, as well as means for deriving these from respective listening tests. Both feature-based and more holistic modeling approaches are analyzed. Overall, open research questions are summarized, deriving trajectories for future work on spatial-audio *Sound Quality* and *Quality of Experience* modeling.

1 Introduction

Sound Quality evaluation¹ has been a research topic since the early days of sound generation and processing, including the evaluation of musical instruments, technical systems such as the telephone, the gramophone or, more recently, audio coding, transmission and large-scale spatial-audio systems. For example, the Bell receiver of 1876, used in the first telephone system, was succeeded by a carbon microphone invented by Edison in 1877 that was reportedly much better sounding than its predecessor—

¹The chapter is a synthesis and extension of the current authors' work presented in Raake and Blauert (2013), Raake and Egger (2014), Raake et al. (2014b), Raake (2016) and Raake and Wierstorf (2016).

Hagen Wierstorf is now with audEERING GmbH.

A. Raake (✉) · H. Wierstorf
Audiovisual Technology Group, Institute for Media Technology, Ilmenau University of
Technology (TU Ilmenau), Ilmenau, Germany
e-mail: alexander.raake@tu-ilmenau.de

see Richards (1973)—*Sound Quality* continues to be the driving forces in the design of audio technology for speech communication or audio systems.

When addressing *Sound Quality*, human listeners are considered who use the received *acoustic* signals to extract features and assign meaning to interact with their environment, in other words, to *communicate* with it. In the audio-technology context of the current chapter, it is assumed that the notion of *Sound Quality* includes any kind of processing between the generation of a sound by its initial source(s) and its recording via different audio-technology systems along the chain up to the listener.

In engineering contexts, instrumental measurements are often used to evaluate and possibly control certain processing steps or technology settings, such as sound pressure levels, frequency responses, decay times, signal delays. They can also include measures related to psychoacoustic features such as intelligibility (Houtgast and Steeneken 1985), or apparent source width (Zacharov et al. 2016b). However, for holistic system evaluation and optimization, *Sound Quality* is addressed as a more integral feature.

Instrumental models of human perception can enable a computational assessment and thus, in principle, in-the-loop control of *Sound Quality*. Such model-based quality optimization has successfully been applied for video-coding in streaming services such as Netflix.² Yet, when designing complex audio technology like spatial-audio or audio-conferencing systems, *automatic Sound Quality* evaluation and respective control mechanisms are still a challenging topic. Hence, especially for newly established reproduction paradigms, listening tests may still be the best-suited approach for some time to come.

This chapter focuses on audio systems at large, and in particular on their *binaural* evaluation, that is, with two ears. A binaural evaluation of *Sound Quality* is generally relevant, and especially when dedicated spatial attributes are evoked by the given auditory scene, such as for spatial audio systems, room acoustics, or the evaluation of sound sources that have a specific spatial extent. Further, certain features also involved in a pure monaural listening may be affected by binaural listening, such as binaural versus monaural loudness (Moore and Glasberg 2007) or binaural de-coloration (Brüggen 2001a). Hence, of particular interest in this chapter are spatial audio systems, where typically both of the above binaural-listening implications are fulfilled. Here, besides monaural also binaural features are involved, evoking respective spatial mechanisms of auditory scene analysis during quality evaluation (e.g., see Raake et al. 2014b).

The concept of *Sound Quality* has been complemented by that of *Quality of Experience* (QoE) during the past 10–15 years. In the literature, the terms *Sound Quality* and *Quality of Experience* are often used interchangeably. However, in the authors' view, *Quality of Experience* represents a more holistic mental construct, related to the entire process of *experience* of a person—see Sect. 2.

As a starting point, the two concepts of *Sound Quality* and *Quality of Experience* are briefly revisited and related to more recent literature. Respective challenges and

²<https://medium.com/netflix-techblog/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f> [last accessed: August 30, 2019].

recent developments in sensory evaluation of spatial-audio systems are discussed. In a subsequent step, the chapter presents a conceptual model of binaural perception, *Sound Quality* and *Quality of Experience* evaluation—see Sect. 5. The description addresses the underlying model concept as well as more concrete aspects for its implementation.³

2 Sound Quality and Quality of Experience

In this section, the concepts of *Sound Quality* and *Quality of Experience* are more formally introduced and set into the context of auditory perception and evaluation.

2.1 Sound Quality

In her work on voice, speech and sound quality, Jekosch defines quality as (Jekosch 2005b, p. 15).

The result of the judgment of the perceived composition of an entity with respect to its desired composition

The underlying concepts are related with the definitions of *Quality of Service* (QoS) by the International Telecommunication Union (ITU-T) and the standardized definition of *Quality* by the International Organization for Standardization (ISO 9000:2000 2000).

In this chapter it is assumed that the definition exclusively addresses perception that “involves sensory processing of external stimuli” (Raake and Egger 2014). Hence, *Sound Quality* addresses the quality evaluation of auditory percepts. In the context of audio-quality evaluation, the term *Basic Audio Quality* (BAQ) is often used for *Sound Quality* (ITU-R BS.1534-3 2015; Thiede et al. 2000; Schoeffler and Herre 2016).

In a technology-related context as in the present book, *Sound Quality* usually addresses a mere technical or technology-related quality, in terms of some sort of *fidelity* or *excellence* (Martens and Martens 2001). In Raake and Egger (2014), a complimentary term, *Assumed Quality*, is proposed as follows.

Assumed Quality is the quality and quality features that users, developers, manufacturers or service providers assume regarding a system, service or product that they intend to be

³The modeling concepts presented are related to the authors’ work in the TWO!EARS project. Evaluating sound quality for *spatial-audio systems* has been one of TWO!EARS’ two proof-of-concept applications (Raake and Blauert 2013; Raake and Wierstorf 2016; Wierstorf et al. 2018). The TWO!EARS-system architecture is open and modular. All documentation, code, data as well as descriptions for hardware implementation are accessible open-source under www.twoears.eu [last accessed: February 18, 2020].

using or will be producing, without, however, grounding these assumptions on an explicit assessment of quality based on experience.

This term was introduced since often the evaluation or even choice of a multimedia technology is made with regard to specifications or physical assessment criteria such as amplitude spectra instead of perception or experience of resulting stimuli—see also the discussion of the *layer model* in Sect. 2.3.

2.2 *Quality of Experience*

The term *Quality of Experience* was introduced to the ICT/multimedia field in the early 2000s as a counterpart to *Quality*, and later standardized by ITU-T in Rec. P.10. An improved definition was developed in the European COST Action *Qualinet* (Qualinet 2012), and was now adopted by the ITU-T in ITU-T Rec. P.10/G.100 (2017). An extended version has been proposed in Raake and Egger (2014). The same definition of *Quality of Experience* underlies the current chapter.

Quality of Experience is the degree of delight or annoyance of a person whose experience involves an application, service, or system. It results from the person's evaluation of the fulfillment of his/or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality, and current state

According to this definition, *Quality of Experience* applies to a judgment of *experience* in terms of “[...] *the individual stream of perceptions, that is, of feelings, sensory percepts, and concepts that occurs in a particular situation of reference*” (Raake and Egger 2014). This definition reflects that the *experience* can have hedonic—that is, pleasure or lack thereof—and pragmatic—that is, concept- or ergonomics-related aspects (Hassenzahl 2001).

The concept of *Quality of Experience* as developed in a multimedia-technology and telecommunications context bears remarkable similarity with the notion of *experienced utility* by Kahneman (1999). According to Kahneman, *experienced utility* refers to a judgment in terms of good/bad of a given experience, related to individually perceived “*pleasure and pain, point[ing] out what we ought to do, as well as determine what we shall do*”—compare Kahneman (2003) with reference to Bentham (1789).

Applied to sound or audio systems, *Quality of Experience* hence reflects the holistic experience of a person when exposed to a scene that contains sound, and in terms of technology assessment, reflects to which extent the integral experience is influenced by the underlying audio technology. Accordingly, it is apparent that *Sound Quality* and *Quality of Experience* are closely related, though not the same. As the next step toward a comprehensive *Sound Quality* and sound-related *Quality of Experience* model, their relation will be analyzed further in view of the *layer model* of Blauert and Jekosch (2012), and Blauert (2013).

Table 1 Quality layers and respective exemplary features applied by listeners for assessment—adapted from Blauert and Jekosch (2012)

#1 Auditive	#2 Aural Scene	#3 Acoustic	#4 Communication
Loudness	Identification	Sound pressure	Product-sound quality
Roughness	Localization	Impulse response	Comprehensibility
Sharpness	Object formation	Transmission function	Usability
Pitch	Intelligibility	Reverberation time	Content quality
Timbre	Perspective	Position	Immersion
Spaciousness	Arrangement	Lateral-energy fraction	Assignment of meaning
	Tonal balance	Cross-correlation	Dialogue quality
	Transparency		

2.3 Layer Model

Blauert and Jekosch proposed a classification scheme of quality according to four different layers of abstraction of the underlying references (Blauert and Jekosch 2012; Blauert 2013), see Table 1, where different features applied for *Sound Quality* evaluation at the different layers are summarized, too.

- #1 The *Auditive* layer addresses *psychoacoustics* references, and relates to fundamental psychoacoustic concepts such as loudness, spectral balance, spaciousness, absence of artifacts. These features do not form aural objects as such but are only components of them.
- #2 The *Aural Scene* layer is related with *perceptual-psychology* references, and refers to the aural-object-formation and scene-analysis step. Instead of analytic listening as for the psychoacoustic features, listeners now focus on object properties and aspects such as their constancy and plausibility (for example in terms of identity). According to Blauert and Jekosch the work of Tonmeisters and sound-engineers is mainly happening at this level (Blauert 2013).
- #3 The *Acoustics* layer incorporates references from *physical acoustics*. It comprises the acoustic-signal analysis, and addresses physical measurements by experts. For these, mathematical abstraction is required. This classification may appear counter-intuitive at first, since this physical level is typically assumed to lie below any other level, that is, based on how persons process acoustically (physically) presented information. The general motivation of including the acoustics level here is that physical descriptors appear to be good correlates of certain perceived features of sound quality.
- #4 The *Communication* layer relates to references from *communication sciences*, in terms of the *meaning* associated with a scene. Here, intra- and inter-personal, cultural and social aspects come into play, and the received *signs* in terms of a semiotic view according to Jekosch are interpreted as a whole (Jekosch 2005b, a). At this level, the process of *experience* is fully involved.

In summary, it can be stated that *Sound Quality* as defined in this chapter encompasses the *Auditive* and *Aural-Scene* layers, that is, #1 and #2. The *Acoustic Layer* #3 is related to the aspect of *Assumed (Sound) Quality* of a system as discussed in Raake and Egger (2014). Accordingly, the *Communication Layer* #4 is related to the concept of *Quality of Experience*.

2.4 Temporal Considerations

For both, *Sound Quality* and *Quality of Experience*, the *time* or *moment* at which the evaluation takes place is relevant (Kahneman 2003; Wältermann 2005). Three time spans are differentiated here, (a) during the experience or *instantaneous*, (b) just after the experience, that is, retrospective judgment on the remembered, as it is often applied in listening tests, and (c) a more episodic view such as retrospective evaluation of a certain event or episode lying further in the past. A corresponding review of the literature can be found in Weiss et al. (2014).

2.5 Influencing Factors

To different extents, *Sound Quality* and *Quality of Experience* depend on a number of influencing factors. According to Reiter et al. (2014), these can coarsely be divided into three main classes, namely, *human*, *system*, and *context*. For this book, *human* is the most important class and is discussed intrinsically in this chapter. The other two classes, *system* and *context*, will mainly be addressed indirectly in the remainder of the chapter, and are briefly discussed in the following.

Context

The perception process and the evoked references depend on the current context of the specific person. The situation is depicted in Fig. 1. The context may influence the role of the acoustic input signals by injecting specific contextual sounds or background noise, into the perception process, by triggering attentional processes, or pre-conditioning peripheral processes, and by steering the expectations in the mind of the listener.

The following example may help illustrate the different levels of contexts and roles. A person is attending a musical performance in a concert hall with friends. The person receives several inputs from different modalities (e.g., auditory and visual) as indicated by the keyword *signal(s)* in Fig. 1. The person interacts with the other persons and, possibly, with the concert hall, for example, by changing his/her position (*interactional context*). The socio-cultural background of the group of friends, who jointly attend the concert, forms the socio-cultural context. How the person under consideration experiences the concert and evaluates *Sound Quality* or *Quality of Experience* depends on the perceived signals and on the further contextual settings. As

such, this information represents the reference-related inputs to the quality-formation process and, thus, to any respective quality model.

Consideration of context also relates to the *relevance of the technology* and hence some underlying, though not consciously addressed aspects of *Quality of Experience*. For example, during a dinner with friends, a certain level of background music may be appreciated, though mainly the type, the specific content, and the loudness of the music will be of relevance for most people. In contrast, during a Jazz concert or in a “high-end”-audio listening situation, the listeners’ attention will be more strongly focused on *Sound Quality* as an important contribution to the overall *Quality of Experience*. Obviously, also the type of listener plays a key role here. An audiophile listener will explicitly include aspects of *Sound Quality* in the overall experience, even more so in a respective listening context—compare the aesthetics-related considerations in Mourjopoulos (2020), this volume.

System

The goal of much of the sound quality-related research is to ultimately understand the impact of technical choices during the implementation and/or configuration of the end-to-end chain—see Fig. 2. This includes all steps from sound recording or capture, mixing, post-production, coding, transmission to presentation (Spors et al. 2013).

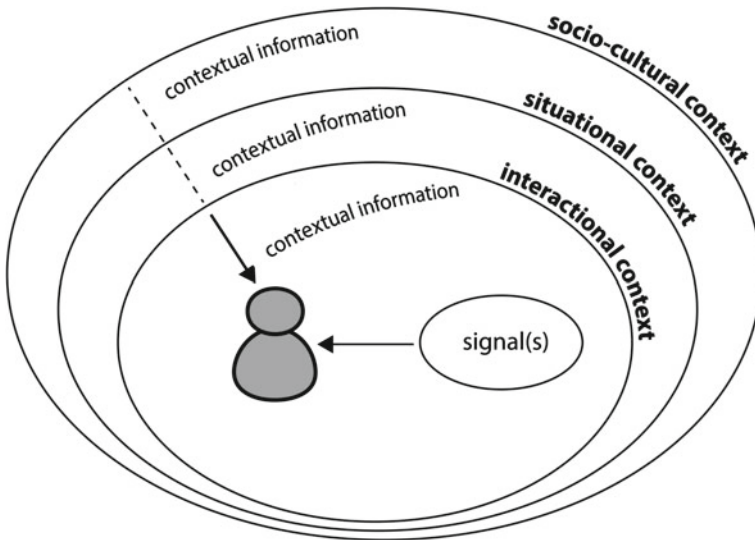


Fig. 1 Contexts of use of an audio-related system (Raake and Egger 2014), adopted from ideas by Geerts et al. (2010) and Moor (2012). The context-dependent roles of the persons, different implications of the physical environment, and of the other actors present in the different types of contexts determine their perception, as well as their evaluation of *Sound Quality* and *Quality of Experience*

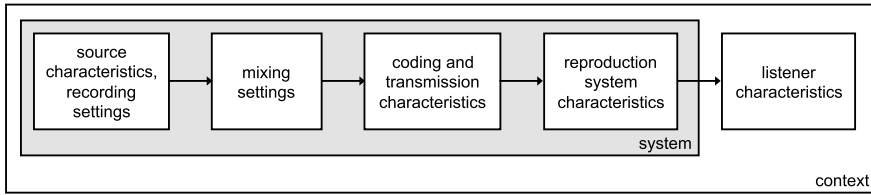


Fig. 2 End-to-end chain (“system”) including sound recording, processing, transmission and reproduction in terms of the factors that ultimately determine *Sound Quality* and *Quality of Experience*—adapted from Wierstorf et al. (2018)

It is important to consider the role of the involved audio technology at all steps. The different characteristics and processing steps (source characteristics, recording, post-production and mixing, transmission, reproduction, and perception) interact with each other, ultimately determining the auditory events. For example, as shown in Wierstorf et al. (2018), the production process cannot be excluded when evaluating *Sound Quality* and *Quality of Experience* of spatial-audio systems.

2.6 Internal References and Expertise

Internal references⁴ in the mind of the listeners are evoked and applied during *Sound Quality* and *Quality of Experience* formation. According to Neisser (1978) and Jekosch (2005b), these are related to the concept of *schema* originating from Piaget’s early work of 1926 (English translation: Piaget 1962). Piaget proposed to consider the schemata-formation processes in terms of *accommodation* (based on revision of internal schemata to include new percepts) and *assimilation* (adjustment of perceptual representation to comply with existing schemata)—Neisser (1978); Jekosch (2005b). These concepts help to understand how internal references are formed—compare Mourjopoulos (2020), this volume. In particular, when listeners encounter types of auditory events that have so far been unknown to them, for example, when listening to high-end spatial-audio systems, enabling 3 D sound, assimilation may happen first by adapting to existing references. Only later, they accommodate to the new perception by learning new references.

Further, it is important to note that *different sets of references* are likely to exist in the listeners’ minds. These depend on different listening contexts, for example, the type of acoustic scene (classical music or an audiobook), the characteristics of the listening room (kitchen or concert hall), and/or the purpose of the listening situation (dedicated listening or a social event).

The formation of internal references are influenced by the degrees of activity and control involved in the reference-built-up, and the intrinsic motivation and interest

⁴The representations available in memory in abstracted form, and used at the different perceptual and evaluation stages.

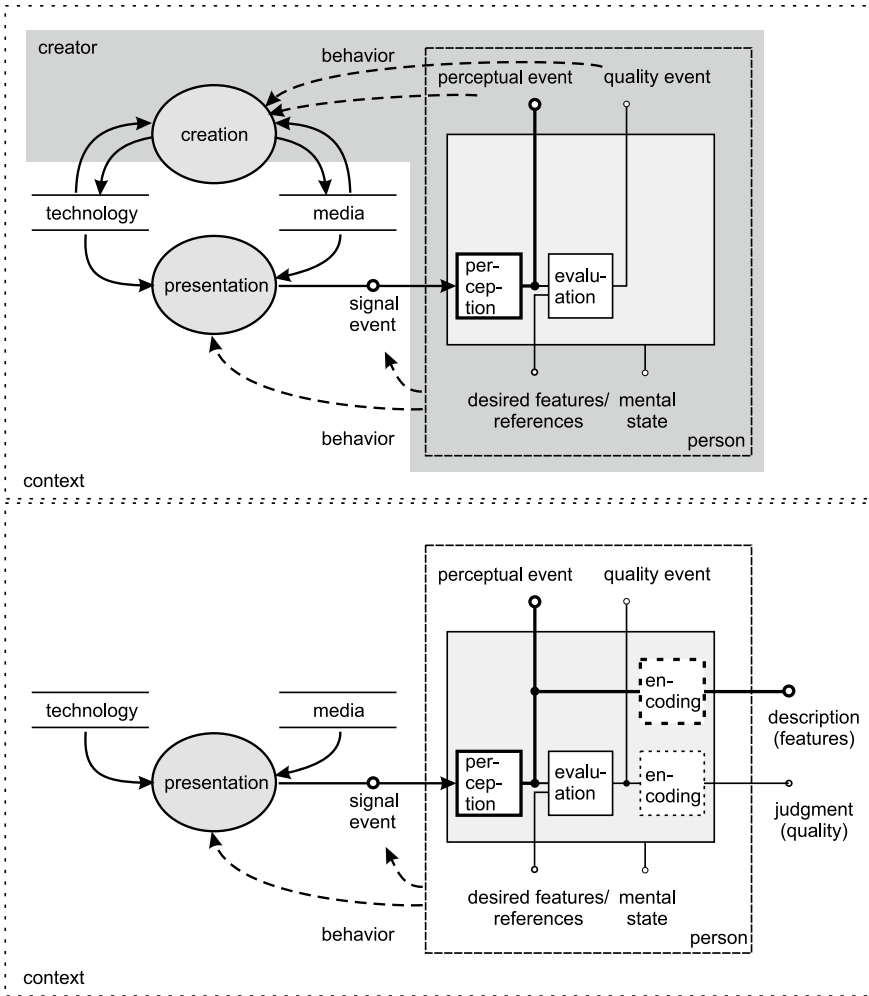


Fig. 3 **Top** Quality perception in the context of creation/production—adapted from Raake and Egger (2014). Perception and evaluation are represented in a simplified manner as two processing components in the mind of a person. The person and creator may be identical and *Quality of Experience* is used as target for optimization. **Bottom** Quality perception during listening. In the case of a listening test and hence *controlled listening*, the impression of *Sound Quality* will be *encoded* into a description or a quality judgment. In the case of *random perception*, without a dedicated listening-test context (Jekosch 2005b), no explicit “encoding” into sound quality judgments or a description will be performed, and both *Sound Quality* and *Quality of Experience* may “happen” in the mind of the persons, depending on the persons’ attention

regarding audio systems. Hence, the reference formation represents aspects of *prior experience* and *expertise*, which is highly related to the work of Kahneman (2011). The different levels of *activity* and *control* during the reference formation can be further specified as

Passive, indirect built-up of references by exposure to different systems during consumption. *Passive* here means that there is no dedicated effort by the listener to control specific system settings or to compare different systems. The rather limited control may be due to little intrinsic interest or expertise regarding the underlying audio technology, or simply due to a lack of opportunity for high-quality-audio listening as a result of lacking availability of cultural resources such as concerts, home-stereo systems, or professional systems. *Indirect* here means that the build-up of references happens indirectly during usage.

Active selections from fixed system options or regarding basic reproduction settings. For example, the listener may be able to make direct comparisons of audio systems in a store or at home and, hence, learn about perceptual differences and own preferences. Further, some degree of control may be available, such as placement of loudspeakers, an adjustable equalizer, or pre-sets that enable modifications of spatial or timbral features.

Active control, where a person may be able to control certain system and media settings so as to realize, based on expertise, the auditory event according to some internal reference, resulting from her/his prior experience. Here, the reference build-up may result from a dedicated training as it explicitly or implicitly happens while learning to play an instrument or to become a professional audio engineer or Tonmeister. It needs to be noted that, in this case, aspects such as talent, type, and quality of training, intrinsic and extrinsic motivation, and availability of technology play important roles for the internal references and achievable degree of control. This highest level of references in terms of the iterative build-up and principal ability to control percepts may be referred to as *realization references*.

The respective process is illustrated in Fig. 3, comparing *listening during creation* that enables substantial control of the source features to *more passive situations*, for example, listening to a recording at home, to a live concert, or as part of a quality test.

3 Sensory Evaluation of Sound Quality and Quality of Experience

A question at this point is, how can *Sound Quality* and *Quality of Experience* actually be assessed. According to Jekosch (2005a), assessment is the “*measurement of system performance with respect to one or more criteria [...], typically used to compare like with like*”, for instance, two alternative implementations of a technology or successive realizations of the same implementation. The judgment criteria can then be certain perceived features or constructs in relation to *Sound Quality* or *Quality of*

Experience. Quality assessment methods can be classified into perception-based or *sensory*⁵ and *instrumental*,⁶ in relation to whether humans or technical systems are used for the assessment (Raake and Egger 2014; Raake 2006).

In the following, the focus will be on tests with human listeners using methods of sensory evaluation. Sensory evaluation is not unique to sound-related quality, but is of relevance in a number of other disciplines such as food quality (e.g. Lawless and Heymann 2010) or service quality in a broader sense (e.g., Parasuraman et al. 1985; Reeves and Bednar 1994)—for more details compare Raake and Egger (2014).

Since *Sound Quality* and *Quality of Experience* are constructs describing certain percepts of humans, sensory evaluation is ultimately the most valid way of assessment. Sensory evaluation tests are usually employed to collect ground-truth data for the development of instrumental methods. For overviews of related test methods see Bech and Zacharov (2006), Raake (2006), and Zacharov (2019).

Sensory evaluation methods can be divided into *direct* and *indirect* ones. With *direct* methods, listeners are directly asked to judge the perceived quality of a presented stimulus or technical system or to rate attributes that characterize the perceived scene or system-related quality impact. Prominent examples of direct sound quality assessment are the methods presented in standards from the International Telecommunication Union (ITU). These include the *Absolute Category Rating* (ACR), applying a rating scale with five or more categories (ITU–T Rec. P.800 1996). The most prominent one of these is the five-point ACR scale, frequently referred to as MOS-scale, where a *Mean Opinion Score* (MOS) is calculated as the average of ratings.⁷ The scale is mostly used for stronger degradation. For intermediate levels of degradation, the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchors) method is recommended, cf. ITU–R BS.1534-3 (2015). For small impairments, “BS-1116” is recommended, cf. ITU–R BS.1116-1 (1997). An overview of the methods recommended for assessing degraded audio is given in ITU–R BS.1283-1 (2003).

Methods that assess constructs related to *Sound Quality* or *Quality of Experience* without the usage of direct scaling or questionnaires are referred to as *indirect* methods. Examples for indirect methods may involve physiological techniques such as measuring skin conductance, heart rate, or EEG—see Engelke et al. (2017) for an overview. Further, behavior-related measures can also be used, based on head motion, facial reactions, task-performance and reaction times.

In short, direct methods guide the subjects’ attention toward the attributes being measured, while indirect methods do not (Pike and Stenzel 2017). The differentiation in terms of direct versus indirect methods is also related to the concepts of *random* versus *controlled* perception (Jekosch 2005b). Random perception refers to perception in natural usage or listening contexts, without an extrinsic test task or laboratory

⁵Often referred to as *subjective*, a somewhat misleading term avoided here.

⁶Often referred to as *objective*, erroneously implying that instrumental measurements bear objectivity, which they only do in case that they can be generalized.

⁷Note that this nomenclature is misleading in at least two ways. First, the ACR scale ideally should be interpreted as an ordinal and not as an interval scale. This means that calculating averages may be inappropriate. Second, any average of ratings may be called “MOS”, that is, not only using the 5-point ACR scale.

environment—see also Fig. 3. In turn, controlled perception occurs for example in a listening test with a concrete listening and judgment task. If done in a way that controlled perception is evoked in the test-listeners' minds, both direct and indirect assessment techniques are likely to yield experimental biases (Zieliński et al. 2008). An alternative is to observe listeners in a non-intrusive manner and to collect behavioral data, such as listening durations, frequency of usage, and actions (for example, play, stop, switch, or head-rotation for visual exploration) and analyze these together with technical characteristics or signals—compare, for example, Raake et al. (2010), Skowronek and Raake (2015), Rummukainen et al. (2018), for audio, and Dobrian et al. (2013), Robitza and Raake (2016), Singla et al. (2017) for video. It should be noted that the *intrusiveness* of the test method is a key aspect. For example, if such indirect assessment is done in a way evoking *controlled* perception—as in laboratory settings where the listeners are aware of the fact that they are in a test situation—the behavior may significantly differ from *random perception* and natural usage—see Robitza and Raake (2016).

3.1 *Sound Quality Versus Quality of Experience Evaluation*

Sound Quality according to its definitions in this chapter reflects the case where the assessors are aware of the technical system or at least the form/carrier (Jekosch 2005a) of the sound and assess it directly. Respective listening scenarios are, for instance, trying out different audio systems for purchase in a store, or taking part in a sound quality listening test. In the case of *Quality of Experience*, the listener is not necessarily aware of the extent to which the listening experience is influenced by the technology used during any of the different steps from recording to reproduction. Due to the associated general difficulty of *Quality of Experience* assessment, most of the literature from the audio-technology domain is restricted to dealing solely with *Sound Quality*.

Assessors listening to sounds that result from the use of some kind of technical system can typically take on two perspectives, namely, (a) focusing on the system that is employed, for example, paying attention to the sound features related to the audio system when reproducing a musical piece or, (b) focusing on the auditory scene or musical piece presented, i.e. on the content. Mausfeld (2003) has described this as the “dual nature” of perception. Research presented in Schoenberg (2016) has underlined the validity of this view when assessing *Quality of Experience* in the context of mediated speech-communication applications. In the case of everyday usage of audio technology, it may happen that degradations due to the technical system are attributed to the audio scene or scene element such as a communication partner (Schoenberg et al. 2014). This often cannot be measured in a test asking for *Sound Quality*, but represents an important contribution to *Quality of Experience* with regard to the overall experience.

Obviously, for music and other types of audio similar considerations apply as for speech communication. For example, in cases where processing steps such as mixing

and reproduction alter the perceptual character of the initially recorded scene, these may be attributed to the scene and not to the involved audio technology (Wierstorf et al. 2018). For example, a singer may be perceived to sing with more passion, when the degree of amplitude compression is increased, or an orchestra may be perceived as spatially smaller or larger when the sound-pressure level is modified.

Hence, assessing *Quality of Experience* relates to the audio experience in a more holistic manner, and implies that the listener is not explicitly aware of the fact that the technology is assessed, thus ideally calling for a more indirect assessment. Respective approaches have addressed preference ratings (Raake and Wierstorf 2016; Wierstorf et al. 2018) or rank-ordering (Rummukainen et al. 2018), the assessment of liking (Schoeffler and Herre 2013; Wilson and Fazenda 2016) overall experience (Schoeffler and Herre 2013), emotional aspects (Lepa et al. 2013), task performance or cognitive load (Skowronek and Raake 2015; Rees-Jones and Murphy 2018), as well as behavioral data collection (Kim et al. 2013).

3.2 *Multidimensional View of Sound Quality*

Sound quality can be assumed to be a multidimensional percept. Hence, a systematic approach to sensory evaluation in terms of multidimensional analysis of perceptual features is appropriate. Such sensory evaluation represents a well-established practice in the food or beverage industry. The totality of perceived *features* describes the *perceived composition*, *perceived nature* or *character* of a sound (respectively Jekosch 2005b, 2004; Letowski 1989).

Specific terminology has been introduced by Jekosch in this regard, distinguishing *quality features* from *quality elements* (Jekosch 2005b). *Quality elements* are, so to speak, the knobs and screws that a designer of the technology, service, or system has at hand to realize a certain level of *Sound Quality* or *Quality of Experience*. *Quality features* are the relevant perceptual features as used by assessors for judging *Sound Quality* or a more integral *Quality of Experience* formation.

The development of multidimensional sensory evaluation methods typically follows several of the steps illustrated in Fig. 4. In the figure, the development of the *sensory measurement system* is illustrated, including both the listening panel and the multidimensional-test and -analysis methods. The upper pathway indicates the sensory evaluation approach with listeners. The lower pathway represents, how the sensory ground-truth data can be used for quality-model development. Here, features for dedicated predictions of quality dimensions and, further, a respective preference mapping to underlying internal references “ideal points” are indicated as an approach to dimension-based quality modeling

The literature on multidimensional analysis of *Sound Quality* includes work on speech quality (Mattila 2002; Wältermann et al. 2010), concert-hall acoustics (Lokki et al. 2011), spatial-audio quality (Rumsey et al. 2008; Wierstorf et al. 2013; Lindau et al. 2014; Zacharov et al. 2016b, a), and audiovisual-quality evaluation (Strohmeier et al. 2010; Olko et al. 2017). In these works, multidimensional analysis

Table 2 Selection of auditory features that are of particular relevance for binaural evaluation. The feature categories are mostly adapted from Zacharov et al. (2016b). They reflect a perceptually motivated rather than a spatial-audio expert-related categorization. For the latter one refer to e.g. Lindau et al. (2014)

Feature	Manner of their specific implication in binaural listening
Loudness	Perceived increase due to binaural listening, Moore and Glasberg (2007)
Coloration	Binaural decoloration using interaural correlation features (Brüggen 2001a, b)
Reverberation	Especially for early reflections, a binaural de-reverberation occurs (Zacharov et al. 2016b; Lindau et al. 2014)
Localization: <i>Distance</i>	Binaural features used in near-field for distance perception (Blauert 1997; Zahorik et al. 2005)
<i>Internality, externalization</i>	Different acoustic, auditory and multimodal effects that determine the amount to which an auditory event is localized either <i>out-of-head</i> or <i>inside-the-head</i> (Hartmann and Wittenberg 1996; Blauert 1997; Brandenburg et al. 2020)
<i>Localizability</i>	Lateral/horizontal-plane localization (Blauert 1997). Related to <i>spatial fidelity</i> and respective modeling approaches as discussed in Rumsey et al. (2008) and Wierstorf et al. (2017a)
<i>Depth, width, envelopment</i>	Interaction between source and playback-room properties in conjunction with binaural hearing (Bradley and Soulodre 1995; Griesinger 1998; Blauert 1997)

techniques such as attribute scaling—with and without prior attribute elicitation—multidimensional scaling, or mixed-methods are used to construct perceptual-feature spaces associated with *Sound Quality*.

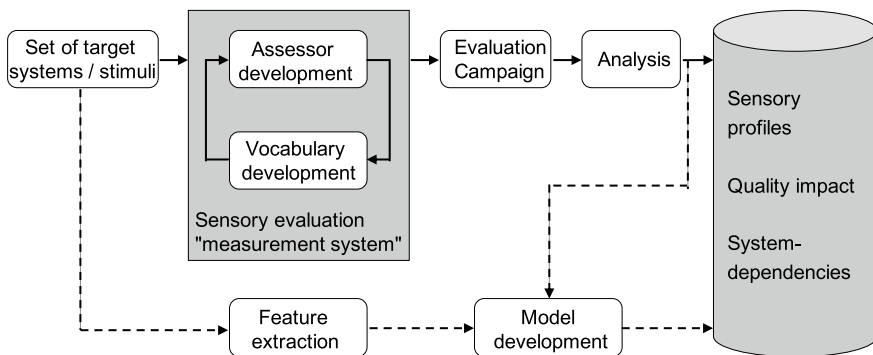


Fig. 4 Steps involved in the development of a sensory-evaluation-test method and, subsequently, of a sound quality model based on multidimensional analysis

Application to Spatial Audio and Sound Quality Modeling

A number of studies on spatial-audio *Sound Quality* have addressed the aspect of *attribute elicitation* (for example, Francombe et al. 2017a; Reardon et al. 2018). It turns out that specific features are particularly important when it comes to *binaural* evaluation of sound quality. A selection of most prominent features in this context is provided in Table 2. Note that all perceptual features are typically affected in the case of a binaural listening versus monaural or diotic listening.

To get from dimensions to *Sound Quality*, (external) preference mapping may be applied (Carroll 1972), relating the multidimensional feature space to uni-scale *Sound Quality*-ratings or preference scores. Frequently, a so-called “ideal point” of the multidimensional feature space can be found that represents the statistically best-possible *Sound Quality* (Mattila 2002; Zacharov et al. 2016b). In principle, the search for an ideal-point marks an implicit way of determining the multidimensional representation associated with the perceptually *ideal reference* in the listeners’ minds.

It is important to note that the features comprised in existing vocabularies are—according to the Layer Model presented earlier—restricted to psychoacoustics (Layer # 1) or perceptual psychology (Layer # 2)—compare Wierstorf et al. (2013), Lindau et al. (2014) and Zacharov et al. (2016a, b). In turn, also features of higher abstraction may be applied to characterize differences between aural presentations. For example, regarding their effect on the *meaning* of a given scene, relevant features are interpreted as scene- or scene-object-related attributes. An example case may be when dynamic compression applied to the voice of a singer alters the timbre in terms of basic psychoacoustic features, yet, it may also alter the perceived nature of the singing voice as initially driven by the creative intent of the singer. Similar effects have been observed during informal listening to some of the stimuli used in Wierstorf et al. (2018), and obviously are heavily used in today’s audio production to create particular aesthetic effects—see also Mourjopoulos (2020), this volume.

Multidimensional analysis of *Sound Quality* represents a viable basis for the implementation of larger-scale quality models. It enables the decomposition of the modeling approach into, (1) a feature-analysis step and, (2) a subsequent preference-mapping, in other words, a quality-integration step—see Fig. 4. For the prediction of quality dimensions, feature-related model components such as proposed in Wältermann (2013), Wierstorf et al. (2014), and Raake and Wierstorf (2016) can be used. A full sound quality model can be realized by appropriate weighting and integration or mapping of the individually predicted features to an integral sound quality estimation score (Mattila 2001; Skowronek et al. 2017).

A related approach was followed by Rumsey et al. (2008), who investigated the intermediate constructs “timbral fidelity” and “spatial fidelity” for loudspeaker-based systems, without an explicit step of multidimensional analysis. Their approach is associated to findings as to which, for stereophonic systems, the variance in sound quality tests is explained to 70% by timbral fidelity and by 30% by spatial fidelity. This holds when no further artifacts such as noise or coding distortions are present

(Rumsey et al. 2005).⁸ Similar findings as those of Rumsey were reported by Schoeffler et al. (2017), confirming a higher contribution of timbral than of spatial effects in MUSHRA-type tests on Basic Audio Quality (note that this depends on the strengths of the related effects initially used in the underlying tests).

3.3 *Spatial Audio Related Challenges*

Obviously, evaluating spatial-audio technology is an application context where *Sound Quality* and *Quality of Experience* are of primary relevance. However, this domain is also intrinsically a quite difficult one for quality assessment. There are a number of particular challenges, for example regarding,

1. The specific system instances under investigation. For example, the perceptual effects resulting from real-life high-quality spatial-audio-reproduction set-ups are rather small compared to degradations due to coding or low-cost electro-acoustic interfaces. Actually, they may be characterized solely by *differences* in particular features but without the system sounding *degraded* at large. As a consequence, test subjects tend to give rather high quality scores overall, or they may not perceive large quality differences even in multi-stimulus comparison tests. Hence, spatial-audio quality can be difficult to assess in perceptual tests and, hence, as well by means of instrumental models trained on respective data.
2. It is likely that there is no established reference in the minds of listeners especially when it comes to commercially rather uncommon spatial-audio-reproduction systems such as massive multi-channel Wave Field Synthesis (WFS) systems. Besides the lack of familiarity with how *spatial* such systems should sound, this may also be due to the lack of an established end-to-end production process for such systems. These effects have more extensively been investigated in the authors' work—described in Wierstorf et al. (2018).
3. Consequently, the scenes most commonly addressed in spatial audio quality tests are rather simplistic and do not play out all advantages and possibilities that such systems may offer. For more realistic scenes, aspects of scene segregation and auditory-scene analysis (Bregman 1990) play a larger role than for simpler audio scenes (Raake et al. 2014b). The creation of appropriate and complex test scenes that can be used for model development is a research task of its own.
4. Further, most recordings and productions with a specific focus on audio have, in the past, addressed pure audio. Today, there is increasing usage of immersive visual-media technology in combination with audio—not only in movie theatres. As a consequence, multimodal interaction plays an even more important role for perception than it does for more traditional, television-like audiovisual content (Garcia et al. 2011). A variety of aspects have to be addressed in testing and,

⁸Note that findings that result in proportionality of relevant perceptual factors depend heavily on the specific test conditions used. Compare Zieliński et al. (2008) for a discussion of biases in listening tests.

hence, also in the underlying scenes, such as, (i) audiovisual attention, (ii) cross-modal feature- and quality-interactions including spatial congruency (e.g., van Ee et al. 2009), or congruency between the visual impression of a room and the perceived room acoustics (Werner et al. 2016; Brandenburg et al. 2020), and congruency of the reaction of a virtual-reality scene with the motion behavior of the viewers/listeners.

From the above challenges it becomes apparent that appropriate test methods are required to collect the ground-truth data for developing models that predict binaural *Sound Quality* or *Quality of Experience* for spatial audio systems, and even more so in the case of dynamic or exploratory, active listening by the users.

3.4 New Approaches for Sensory Evaluation of Spatial Audio

Even for more traditional scenes that enable 3-DoF motion, asking for some sort of *quality* represents a challenging task for test subjects. Besides scaling-related biases described by Zieliński et al. (2008) and others, recent work on direct rating has shown a bias towards *timbral* (audio) or *signal-clarity* (video) features—see, for instance, Zacharov et al. (2016b), Benoit et al. (2008), and Lebreton et al. (2013) with a similar study for video quality. This refers to the notion of *excellence* of sound quality as discussed in Sect. 2.

Still, the widely used MUSHRA-type ratings of sound quality can be considered as a viable approach as long as the focus is clearly restrained to *Sound Quality* or *basic audio quality* (ITU-R BS.1534-3 2015). Related to the difficulty and often absence of a dedicated reference in the context of spatial audio quality assessment, reference-free variants of MUSHRA are clearly preferred in this context. Whether both expert and less experienced listeners can validly *directly* rate a more holistic *overall listening experience* using such a MUSHRA-type approach remains questionable to the authors of the current chapter—compare Woodcock et al. (2018). A corresponding, MUSHRA-based method that addresses the relation between basic-audio-quality scores and underlying attribute ratings from experts has been presented by Zacharov et al. (2016b).

Paired Comparison

As an alternative to implicitly fidelity-focused, direct methods, comparative methods such as Paired-Comparison (PC) preference tests can be used. They help to avoid some of the possible biases and address the challenges of a generally high quality and hence restricted sound quality range, and work also for complex scenes. Examples of related work are Wickelmaier et al. (2009), Li et al. (2012), Lebreton et al. (2013), and Wierstorf et al. (2018). In a PC-type preference test, listeners are asked to rate which presentation or version of a given pair of stimuli or systems they prefer. Hence, the binary rating task for the listeners is a rather simple one.

A respective approach in-between *Sound Quality* and *Quality of Experience* assessment has been taken in more recent work by the authors for spatial-audio eval-

uation (Wierstorf et al. 2018). With such a PC-based approach, it can be assumed that both technology-oriented and more hedonic aspects, related to the “meaning” of the audio piece after interaction with technology, are used by the assessors when deciding for preference. This is particularly the case for non-expert listeners who have no deeper knowledge of the processing applied. The PC-test paradigm was used for assessing preferences between pairs of spatial-audio processing and reproduction conditions, including different combinations of mixing/post-production and spatial-audio presentation. Three spatial-audio-reproduction methods were compared, namely, stereo, 5.1 surround sound, and wave-field synthesis (WFS), with different variations of sound mixes produced specifically for each reproduction method. In all three tests, WFS turned out to be the most-preferred reproduction method. However, the amount of preference was reduced to a large extent when a less preferred mix was applied, almost neutralizing the reproduction-related advantage.

In Francombe et al. (2017a, b), a combination of sensory evaluation in terms of attribute ratings with paired-comparison preference is reported. Different audio excerpts were presented, using a number of spatial-audio reproduction methods, namely, headphones, “low-quality mono”—that is, small computer loudspeakers, further, mono, stereo, 5-channel, 9-channel, 22-channel, and ambisonic cuboid. Both experienced and inexperienced listeners were recruited as assessors. Different sets of attribute vocabulary and scales were developed, one for each listener group. The attribute elicitation was part of a pair-wise preference test. By comparing across all stimuli, a preference for 9- and 5-channel over 22-channel was found. However, considering the results of Wierstorf et al. (2018), it remains questionable in how far less adequate mixes or the underlying source material as such may have lead to the lower preference for the 22-channel reproduction.

A further indirect alternative to fully-paired comparison is *rank ordering*. Such tests aim at reducing the number of comparisons by iteratively eliminating the least favored of different stimuli based on an intrinsically reduced number of paired comparisons (Wickelmaier et al. 2009; Rummukainen et al. 2018). These methods have been shown to reduce the required testing time in comparison to PC-tests with full pairs. Rummukainen et al. (2018) conducted a preference-based ranking test for different audio scenes both in an audiovisual VR and offline, evaluating the contribution of different audio-rendering methods. In addition to the rank-ordering-test results, different types of behavioral data were recorded, including 3-DoF head rotations—that is, yaw, pitch, and roll. The rank-order results revealed significant differences between audio-rendering methods. The behavioral data, in turn, did not provide additional insight for the system comparisons.

Liking and Sound Quality

In both rank-ordering and direct paired-comparison tests, different versions of the same audio piece are typically compared with each other, that is, differently mixed, processed and/or rendered variants. However, other approaches for more QoE-related assessment—also comparing different audio pieces—have addressed the judgment of *liking* (Wilson and Fazenda 2016; Schoeffler and Herre 2013) or of overall experience

(Schoeffler and Herre 2013, 2016; Schoeffler et al. 2017), as well as possible relations to *Sound Quality* (i.e. Basic Audio Quality).

Schoeffler and Herre (2016) and Schoeffler et al. (2017) have conducted a number of test runs with expert and non-expert listeners, using the following approach. In a first session, the *liking* of individual pieces from a larger number of down-mixed stereo sequences from different genres is judged, later referred to as the “basic-item rating”.⁹ In a subsequent set of sessions, liking is assessed with the same approach for a number of processed (such as band-pass filtered) and differently presented (such as different spatial-audio techniques) sequences subsampled from the initial set of sources. The resulting ratings are referred to as *Overall Listening Experience* (OLE). As the last step, the *Basic Audio Quality* (BAQ) is assessed using the MUSHRA technique (ITU-R BS.1534-3 2015). The obtained data indicate that OLE ratings result from different weightings of the “basic item rating” (*liking*) of the pieces by an individual assessor, and of the BAQ. While this approach represents a novel approach to assess some cognitive constructs closer to QoE than in most other tests, some systematic aspects may raise the question of how close this approach really comes to it. In particular, directly asking for ratings after different presentations and the small number of but still present repetitions of the same contents may cause a higher focus on BAQ than on what really affects the QoE in cases of *random-perception* (Jekosch 2005b) as under real-life listening conditions.

In another study by Wilson and Fazenda (2016), it was hypothesized that *Sound Quality* and *liking* represent independent concepts, with *Sound Quality* referring to a *pragmatic* and *liking* to a *hedonic* construct within the minds of listeners. However, since the listeners were presented with *liking* and *Sound Quality* rating scales in the same test run, the independence of the two rating results may also stem from a test-inherent bias, where subjects may have intended to “de-correlate” their usage of the scales—see also the considerations in Raake and Wierstorf (2016).

Emotional Aspects

A further way to approach *Quality of Experience* may be to assess emotional aspects related to audio listening. For example, Lepa et al. (2013, 2014) conducted tests on emotional expressiveness of music for pieces available both commercially on CD and as multi-track versions. The pieces were processed and played back with three different types of spatialization, using dynamic binaural re-synthesis for presentation, namely, (1) the original CD stereo version, (2) a stereo-loudspeaker simulation using binaural room impulse responses (BRIRs) and, (3) a simulated live event with respective placement of sources on some virtual stage. The listeners judged aspects of the emotional expressiveness for one of the three presentation types using a between-subject design. At the end of each trial, they gave ratings of sound quality attributes using a semantic differential. It was found that spaciousness had a significant effect on the emotional attributes ascribed to the musical performance. In turn, only the sound quality attributes directly related to spaciousness were affected by the presentation type. Lepa et al. (2014) argue that the increased feeling of being surrounded

⁹It may be argued that the specific down-mix may have affected the liking already, depending on the piece and its original recording. It is difficult, though, to address this topic in a different way.

by the sources in the case of a higher degree of spaciousness may be the reason for perceiving a stronger emotional expressiveness. The finding that the three presentation types only affected spaciousness-related sound quality attributes can likely be explained with the fact that the processing mainly differed in spaciousness-related technical characteristics. The proposed approach can be considered as an interesting step towards more QoE-type assessment. However, further research is required to assess how different types of audio processing and presentation may affect not only the perceived emotional expressiveness (i.e. related to musical intent) but also with regard to the emotional state of the listeners.

Behavior and Physiological Assessment

Another indirect approach for evaluating spatial-audio technology includes the assessment of listening behavior or respective task performance. For example, Rummukainen et al. (2017) investigated the performance of persons in a 6-DoF navigation tasks in a VR environment for three different types of spatialization of the sound sources used as targets of the navigation action (Rummukainen et al. 2017). In this pilot experiment it was found that monaural presentation with intensity rendering lead to significantly worse performance as compared to binaural presentation with and without 3-dimensional rendering. Four performance measures were used, that is, mean time to target, mean path length to target, error at the end, and aggregate rotation angle applied. In a further experiment by Rummukainen et al., besides MUSHRA-type reference-free ratings, also head-rotation-behavior data were collected for different binaural rendering engines. The experiment used 6-DoF-VR interactive audio presentation (Rummukainen et al. 2018). While the quality ratings were well indicative of the advantage of individual rendering algorithms, the collected behavior data did not provide any additional information on quality.

Further examples for behavior- or, better, performance-related assessment of spatial versus non-spatial audio are discussed in Rees-Jones and Murphy (2018). One of the studies addressed the impact of spatial audio on the success of players in an audio game. The general idea behind this study was in line with other work on performance in VR-type environments—compare the work reviewed in Bowman and McMahan (2007). However, the game used was very specific with regard to assessing the value of audio. Hence, a transfer to more real-life game usage with complex scenes and a gaming-situation-specific musical-score generation cannot readily be made.

In addition to perceptual and behavioral data, physiological signals can be employed for quality evaluation. In the context of quality or QoE assessment, physiological methods and measures so far employed have been pupillometry, heart rate, skin conductance, brain imaging, EEG (electroencephalogram) including ERPs (event-related potentials), MMN (mismatch negativity), and oscillation analysis—see Engelke et al. (2017). Physiological measurements principally enable *indirect* assessment of latent reactions. This is a suitable approach, especially when these reactions cannot easily be controlled by the test listeners—such as certain emotional responses. Up to now, physiological measurements cannot fully replace perception- and/or behavior-scaling methods since physiological correlates of quality must still

Table 3 Selection of perceptual studies on spatial-audio *Sound Quality* or *Quality of Experience* and availability of data

Study	Data collected	Available?
Choisel and Wickelmaier (2007)	Attributes, preference	No
Zacharov et al. (2016b)	Attributes, quality	No
Reardon et al. (2018)	Attributes, preference	No
Woodcock et al. (2018)	Experience	Unclear
Francombe et al. (2017a, b)	Attributes, preference	Unclear
Raake and Wierstorf (2016)	Localization, head-rotation, coloration, preference	Yes
Wierstorf et al. (2018)	Preference	Partly
Schoeffler and Herre (2016)	Quality, listening experience	No
Schoeffler et al. (2017)	Quality, listening experience	No
Wilson and Fazenda (2016)	Sound quality, liking	No
Lepa et al. (2013, 2014)	Emotional attributes	No
Rees-Jones and Murphy (2018)	Attributes, quality, performance	No
Kim et al. (2013)	Head-motion	No
Rummukainen et al. (2017)	Localization, head-motion, performance	No
Rummukainen et al. (2018)	Quality, head-rotation	No

be related to direct quantitative analysis. For the case of speech quality, this link has recently been investigated in Uhrig et al. (2017, 2018).

3.5 Data Availability and Reproducible Research

A main limitation for model development is the lack of available test material that can be used for training the models. To be clear, this is not only a problem due to a lack of appropriate test methods or the difficulty of running such tests. In addition and possibly even worse, the majority of existing test data has not yet been made available to the research community. In particular, in the domain of sound quality assessment, the currently debated issues of *reproducible research* and *open science* are well behind their potential (Spors et al. 2017). Particularly for *Sound Quality* and *Quality of Experience* research, only little data have been made publicly available. An example of reproducible research is the TWO!EARS project, where most of the results and data are freely available—see, for instance, Wierstorf et al. (2017b, 2018) and Winter et al. (2017). Different studies referenced in the current chapter and the possible usage of their test data for modeling are summarized in Table 3.

4 Instrumental Evaluation of Sound Quality and Quality of Experience

Once appropriate ground-truth data are available, actual model development can be addressed. In this section, different existing models will briefly be reviewed in relation to the model outlined in Sect. 5. Raake et al. (2014b) distinguished two fundamental types of methods in this context, namely,

1. Algorithms or metrics that are based on physical properties of the signal or sound field, which may be put into relation with perceptual attributes or ratings.
2. Algorithms that implement specific parts of human auditory signal processing, possibly including cognition-type mapping to quality dimensions, *Sound Quality* or *Quality of Experience*.

An example of measures of Type 1 for the case of sound field synthesis is a quantitative descriptor to characterize the deviation of the reproduced sound field from the desired one (Wierstorf 2014). An example for room acoustics evaluation metrics are reverberation-decay times (Kuttruff 2016). Such direct relation with physical properties of the sound field may principally enable a more diagnostic control or optimization based on system settings. However, respective measures do not well capture the ground-truth data from sensory evaluation, resulting from human perception and judgment, and certainly do not meet the criteria put forward for the conceptual model proposed in Sect. 5.

To this aim, the *explicit modeling* of human signal processing—see Type-2 measures above—and mapping of perceptual features to sensory evaluation results has to be performed. Various notable approaches of this type have been developed in the past years, and have been standardized in bodies such as the International Telecommunication Union. Examples include *Perceptual Evaluation of Speech Quality* (PESQ) (ITU–T Rec. P.862 2001) and *Perceptual Objective Listening Quality Analysis* (POLQA) (ITU–T Rec. P.863 2011; Beerends et al. 2013) for assessing the quality of speech transmission systems, and *Perceptual Evaluation of Audio Quality* (PEAQ) (Thiede et al. 2000) for audio coding evaluation. Such signal-based, *full-reference* (FR) models estimate quality by comparing the processed audio signal with an unprocessed reference, on the basis of a transformation of both signals into perceptual representations using models of human audition. Further examples of FR-type *Sound Quality* models for non-spatial audio have been presented in Harlander et al. (2014), Biberger and Ewert (2016) and Biberger et al. (2018).

For the instrumental assessment of loudspeaker-based sound reproduction, initial models were constructed on the basis of the notion of spatial and timbral *fidelity* (Rumsey et al. 2005). To this aim, underlying technical or physical characteristics of the acoustic scene were mapped to low-level attributes or perceptive constructs. In the respective model named *Quality Evaluation of Spatial Transmission and Reproduction Using an Artificial Listener* (QESTRAL) (Rumsey et al. 2008), spatial fidelity is predicted from perceptually relevant cues such as interaural time and level differences. Some approaches have been proposed for timbral-fidelity prediction, too.

Moore and Tan (2004) describe a model for coloration prediction of bandpass-filtered speech and audio. Another coloration-prediction model for room acoustics is presented in Brügger (2001b), and a simple speech-coloration model in Raake (2006). For spatial audio, a model based on Moore and Tan (2004) has been implemented within the TWO!EARS framework (Raake and Wierstorf 2016).

As stated earlier in this chapter, Section 3.2, the approach of modeling *Sound Quality* on the basis of individual quality dimensions is generalizable. Starting from relevant predictors of individual quality dimensions, a kind of *external preference mapping* can be applied and the *Sound Quality* can be predicted based on the individual dimensions (Mattila 2001; Wältermann 2013; Choisel and Wickelmaier 2007).

Full-reference models for spatial audio reproduction were under development in ITU-R SG6 (Liebetrau et al. 2010), and different algorithms have more recently been described in the literature (Seo et al. 2013; Härmä et al. 2014). Full-reference models that deal with audio coding analyze first the processed and reference signals in terms of *Model Output Variables* (MOVs), for example, by models of the auditory periphery. In subsequent steps, aspects of human cognition are applied, for example, targeting a relevance-weighting of different MOVs (Thiede et al. 2000; Seo et al. 2013; Härmä et al. 2014).

The different modeling approaches presented up to now account only for some of the targeted capabilities of the conceptual model presented in this chapter. In particular, building up a representation of the world knowledge of listeners is a complex problem. The team behind PEAQ have considered this problem (Thiede et al. 2000), indicating that an explicit reference as in the case of such a *full-reference model* is suboptimal, since, for example, a given processing may improve the signal over the reference. Instead, the “ideal audio signal [...] in the mind of the listener” should be known.

The handling of the problem of an explicit versus internal reference has been addressed in the full-reference, speech-quality model POLQA (ITU-T Rec. P.863 2011). As a new way ahead, it uses an *idealization* step when processing the reference signal, with the following two goals. (1) It reduces different types of non-speech distortions before loudness spectra are calculated in the perceptual model. These are later addressed in a separate processing step for both the reference and the transmitted speech. Interestingly, this approach may be related to a kind of feature constancy targeted by human auditory peripheral processing. (2) Using idealization, sub-optimal reference signals that may be affected by noise or reverberation are transformed into an improved version and thus better representation of the assumed internal reference. This approach addresses the limitations of a fixed reference as the first step towards an actual learning of internal references.

Another topic to be addressed with regard to sound quality models—especially for spatial audio—is the aspect of scene analysis and respective adaptation of the evaluation to specific objects in a scene. The need for a scene-specific evaluation scheme has been addressed in Raake et al. (2014b). For non-spatial audio this issue has been mentioned in Thiede et al. (2000), indicating that certain spectral-temporal artifacts may be processed as distinct streams by listeners and hence may require dedicated stream segregation. The first implementation of a simple scene-analysis model for spatial

fidelity was proposed in Rumsey et al. (2008), using some foreground-background separation following the respective framework for scene-related evaluation as suggested in Rumsey (2002).

In summary, it can be said that to date none of the available approaches comes close to the conceptual model that will be outlined in the subsequent section.

5 A Proposal for a Conceptual Sound Quality Model

In the following, the basic architecture of an instrumental *Sound Quality* and *Quality of Experience* model is outlined. It provides an updated view on the modeling concepts described in Raake and Blauert (2013) and Raake and Egger (2014), based on work of the interest group *Aural Assessment By means of Binaural Algorithms* (AABBA) (Blauert et al. 2009) and the TWO!EARS projects, following the lines of thinking also discussed in Blauert et al. (2013). The model can be considered as a hybrid between, (a) the authors' view of *Sound Quality* evaluation and *Quality of Experience* formation as it occurs in a person's mind, as described earlier in this Chapter and in (Raake and Egger 2014) and, (b) as proposed implementation of certain functional processes of perception and cognition as outlined in Raake and Blauert (2013).

5.1 Model Overview

The model represents a listener who interactively explores the environment based on binaural information, with some crossmodal information considered, too. The model architecture is depicted in Fig. 5. Some of the functions and processes of human perception and cognition are represented by blocks, according to a technical, block-diagram-type processing perspective. For these components, rough concepts or actual implementations do already exist, for example, in the TWO!EARS model framework,¹⁰ or as part of a number of other existing auditory-perception models and toolboxes. Some types of *memory* or information stores and functional processes are outlined as semi-transparent-surface blocks, highlighting that their inclusion into an actually implemented technical model requires further research.

The non-auditory information as considered in the figure primarily addresses the visual sense. As illustrated, *Sound Quality* and *Quality of Experience* evaluation involve high-level cognitive processes, such as psychological and state-related processes like memory, motivation, emotions, and cognitive reasoning. The model combines bottom-up signal-driven processing with top-down hypothesis processing (Blauert and Brown 2020, this volume), for feedback processes involved. The listener interacts with a scene that is represented by multimodal signals as input to the human sensory organs (Raake and Egger 2014). The sensory organs perform a transformation of the physical input signals into neural representations that include

¹⁰www.twoears.eu [last accessed, August 30, 2019].

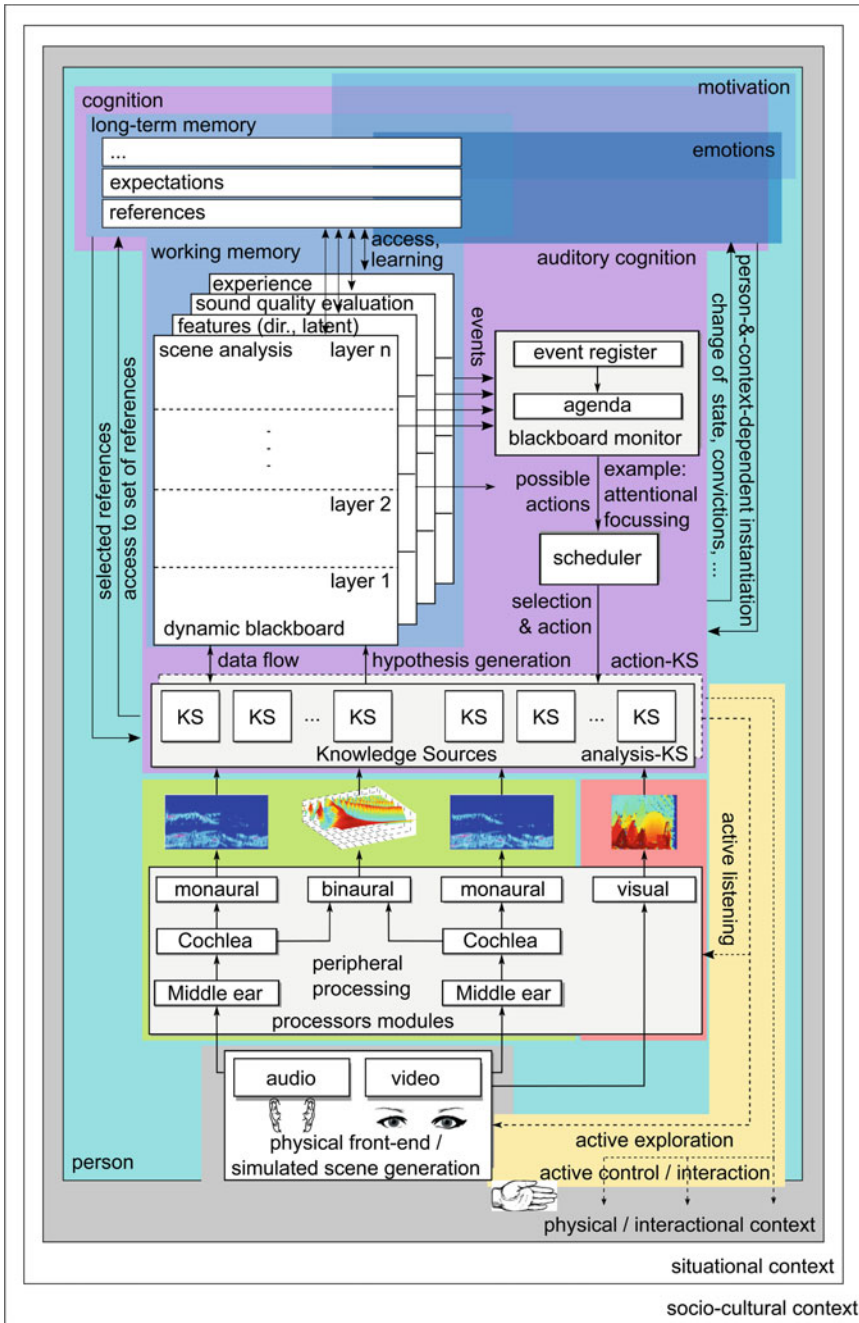


Fig. 5 Architecture of a comprehensive model of auditory and multimodal perception, *Sound Quality* and *Quality of Experience* formation. The picture is based on a conceptual drawing of a specifically tailored blackboard system (Raake and Blauert 2013), later amended by G. J. Brown and N. Ma in the course of the TWO!EARS project (Brown et al. 2014)

characteristic electric signals. The lower-level sensory representation is processed further along the neural pathways to higher brain levels where more abstract, symbolic representations are built (Brown et al. 2014).

It is obvious from the previous discussions in this chapter that dynamic auditory-scene analysis forms the basis for *Sound Quality* and *Quality of Experience* formation. This involves an analysis of and possible adaptation to aspects of the room geometry and the amount of reverberation, to the spatial positions, movements and spatial extents of sound sources, the source identity and further auditory-event attributes, and the assignment of meaning to speech and other types of sounds. Accordingly, the objects of a perceived scene are characterized by different auditory and related crossmodal features (Raake and Blauert 2013; Raake et al. 2014a). These features together form the aural or multimodal character of the objects and scene at large. The mental scene representation in terms of recognized objects of perception is established as an interleaved process of top-down hypothesis generation and their verification against bottom-up perceptual evidence (Blauert et al. 2013; Blauert and Brown 2020).

As a precursor for object formation and scene analysis, the *peripheral processing* delivers a multidimensional, topologically organized representation of the scene, covering aspects of time, space, frequency, and activity (Raake and Blauert 2013; Blauert et al. 2013; Raake et al. 2014a). The neural representation comprises auditory and associated multimodal cues, such as on-/offsets, amplitude modulation, periodicity, interaural time and level differences (ITDs, ILDs) across frequency bands, and interaural coherence, all including their respective timing information, illustrated in Fig. 5 by different spectrogram-type pictograms. The neural representation is assumed to precede the actual formation of perceptual objects (Raake and Egger 2014). The involved steps are performed at higher level by parallel and intertwined processors addressing the bottom-up pre-segmentation of the multidimensional feature representation, essentially carrying out a *Gestalt*-related analysis (this volume, Sotujo et al. 2020). The pre-segmented representation is further analyzed in terms of objects in the specific modalities, such as visual objects or aural scene objects, or words in an utterance.

Various kinds of memory are involved at all of these processing stages. For example, certain representations may evoke remembered perceptual events and subsequent feedback-based adaptation of the processing, such as, for example, noise suppression once a human voice is sensed. At this stage, information from other modalities is already integrated. The inclusion of top-down feedback paths reflects human mental processing of sensory information, beyond the more traditional, bottom-up view of auditory perception. In an implementation—see Sect. 5.2—their start- and end-points at different levels of the model structure need to be specified, as well as the type of information/action that is communicated to the respective lower level(s). Such feedback mechanisms include attention, comprising a selection of bottom-up features, or commands such as exploratory head movements.

Not only the direct sensory signals that characterize a scene are processed by the sensory organs. They also process the contextual and/or task-related information given to a person. Contextual information either directly affects the perceptual

process or does so via evoked higher-level concepts. By definition, perception is determined by the person's current state, that is, "*situational or temporal changes in feelings, thinking, or behavior [...]*" (Amelang et al. 2006, translated from German).

Memory and Perceptual References

In Fig. 5, different parts of memory are illustrated. Research on human memory has identified different levels, with respective roles in the perception process, and respective storage durations.

Sensory memory Peripheral memory, stores sensory stimulus representations for short durations between 150 ms and 2 s, made available to higher processing stages. For auditory information, this storage is referred to as *echoic memory* with storage durations between 150 and 350 ms. For visual information, it is referred to as *iconic memory* with up to 1-s storage duration (Massaro 1975; Cowan 1984; Baddeley 1997; Coltheart 1980).

Working memory Re-coded information at symbolic level for longer durations from a few up to tens of seconds. It is assumed that there are three main storage components involved with working memory, namely, the *visuospatial sketchpad*, the *episodic buffer*, and the *phonological loop* (Baddeley 2003).

Long-term memory Covers longer time spans up to years or even a full lifetime, involving multiple stages of encodings in terms of symbolic and perceptual representations. Current theories assume that a central executive component controls the linking between long-term memory and working memory via an episodic buffer at working-memory level that integrates information into episodes, and that this central component is associated with attention (Baddeley 1997, 2003).

Internal, *perceptual* references in the mind of a listener are assumed to be present at or made available to different levels of memory, namely, in the working memory for the perceptual integration of a scene and respective scene analysis, as well as information in the form being retrieved from long-term memory, for example, for the identification of objects in a scene or words in an utterance. Similarly, the *perceived character* or the respective perceptual event or flow of events can be situated in working memory, and/or be stored in long-term memory, for example, after verbal or episodic re-coding has occurred as the result of a learning process (Raake and Egger 2014). Complementary considerations on categories of references can be found in Neisser's cognitive system theory (Neisser 1994). Neisser assumes that learning is implicitly integrated into perceptual processes and related to aspects such as expertise and know-how with the specific percepts.

5.2 Considerations Regarding Model Implementation

System-type implementations of the conceptual model require a multi-layered architecture with various different modules for bottom-up as well as top-down processing

during interactive exploration—see Fig. 5. Further, the different layers implicitly represent storage components in the system and are interconnected with long-term memory related to the interleaved steps of *scene analysis*, the formation of *Sound Quality*- or *Quality of Experience features* of individual perceptual objects as well as for the scene at large—for instance underlying episodic statements such as, “The singer’s voice sounds great”, “The guitar sounds bad” or, “This is really a very nice concert”.

The listeners integrate various of the lower-level results in light of their current emotional and cognitive state. Events at any of the processing levels may result in intentions-for-action. For example, to re-listen to a certain passage of a stimulus in an audio test, the listeners may press a replay button, change position to achieve a better sound quality, or simply focus their attention on certain aspects of the presented audio material. During learning from present and past episodes, the aural character is transformed into internal references as part of long-term memory. The telephone or stereo systems are examples where most listeners already have an internal reference (Jekosch 2005b).

All perception and subsequent evaluation is done in relation to such internal references—see also Sect. 2. Sets of reference features, *references*, are evoked by the listeners *expectations* in a given *listening context*, and are related to perceived features, for example, triggered by a sound quality evaluation task in a listening test, or when listening to different Hi-Fi systems as part of a purchasing decision in a shop. See *references* and *expectations* in the top left part of Fig. 5, and *features* underlying the sound quality evaluation as listed in Table 2.

While some of the perceived features may directly be nameable by a person—*direct features*, indicated as “dir.” in Fig. 5—for some other features this may not be the case and, hence, direct assessment will not be possible or at least quite difficult for these.

Any implementation of such a complex model system will benefit from a modular software architecture. In its most complete form, two types of realization of such a model are conceivable, (i) a virtual agent that actively explores a virtual scene in software, see for example the related work in TWO!EARS, cf. Blauert (2020), for the assessment of a spatial audio system during the design phase for different (virtual) rooms, and (ii) a model being built into a physical robot system that enables usage with real-life acoustic scenes, including human-like head-and-body displacements within the scene.

The list below summarizes the respective modules, which are briefly outlined in the following.

1. Physical front-end or acoustical simulation that provides ear signals.
2. Binaural bottom-up signal processing that extracts low level features.
3. Pre-segmentation based on low level features.
4. Cognitive processes to build hypothesis on the perceived scene.
5. Feedback mechanisms that can influence all underlying modules.

Front-End and Acoustic Signal Processing

The bottom layer represents the physical front-end of a person or the model system and the respective acoustic and multimodal signal processing involved during the capture of sensory information about the scene. In the case of a human listener, this front-end comprises two ears, head, and body as well as the subcortical, hence peripheral auditory processing. For implementation purposes, the system may have a real physical front-end such as the robotic system developed in TWO!EARS.¹¹ While the TWO!EARS physical-system implementation enabled 3-DoF motion (1-DoF head panning, 2-DoF lateral displacement), real-life interaction of a person with a scene provides a 6-DoF-perspective, namely, 3D displacement in space as well as all three axes of possible head turning (pitch, yaw, roll). With the latest developments of audio reproduction systems with sound-field synthesis such as WFS or binaural-re-synthesis, for example, for Virtual Reality (VR) applications with Head-Mounted Displays (HMDs), 6-DoF has become a highly relevant topic, also with regard to *Sound Quality* and *Quality of Experience* assessment. Alternatively, a virtual system may be employed so as to assess quality based on recorded or synthetically created acoustic and possibly multimodal scenes. As for real-life, interactive binaural listening using loudspeaker set-ups or binaural re-synthesis, respective sound fields or binaural signals must be generated as model input so as to correctly represent the acoustic scenes at the listeners' two ears. For an interactive implementation, head-position information needs to be provided from the model to the scene-generation module to generate the appropriate aural signals.

Auditory Periphery and Pre-segmentation

The subsequent layer addresses the monaural and binaural subcortical bottom-up processing. The input is the binaural ear signals from the bottom layer, representing different scenes with multiple active sources. From this information, primary cues are extracted, (a) monaural cues, including onsets, offsets, amplitude modulation, periodicity, across-channel synchrony, and others, (b) binaural cues, including interaural time and level differences (ITDs, ILDs) across frequency bands, interaural coherence (IC), and others.

Based on these cues, the pre-segmentation can be carried out. Here, features for identification of active sources will be identified, to enable, for example, localization, speech activity recognition, and the source identification. The output of this stage is a multidimensional auditory representation in terms of *activity maps*. These are organized in a topological manner, for example, in terms of time, frequency, and activity. Based on this multidimensional representation, features for auditory scene analysis are extracted, for instance, features temporally collocated across different spectral bands. Moreover, for a sound quality- or QoE-model, respective dedicated features or variations of the psychoacoustics and aural-scene-related features can be extracted.

¹¹Incorporating a head-and-torso-simulator (Kemar) with a motorized neck to enable horizontal-plane panning, mounted on a carriage for lateral motion. See <http://docs.twoears.eu/en/latest/> [last accessed February 22, 2020].

In an actual model implementation, lower-level peripheral processing could be implemented as a collection of processor modules, as has been done with the *Auditory Front End* (AFE) in the TWO!EARS project.¹² In a complete model, these processors can be adjusted by feedback from higher model levels during run time. Feedback could, for instance, lead to on-the-fly changes in parameter values of peripheral modules, like the filter bandwidths of the basilar-membrane filters. To this aim, an object-oriented framework is required, for example, to allow for direct switching between alternative modules while keeping all other components unchanged. Further, for an instantaneous evaluation of *Sound Quality* or *Quality of Experience*, online processing of the two-channel ear signals is needed. In this way, different temporal aspects of quality evaluation can be addressed—compare Sect. 2.2. The cues may also represent the basis for quality integration based on estimates of underlying quality dimensions. In previous work by the authors' group, for example, the cues available from the TWO!EARS project were shown to enable the estimation of localization and coloration, as well as estimation of preferences between stimuli pairs (Wierstorf et al. 2017a, 2014; Raake and Wierstorf 2016; Skowronek et al. 2017).

Cognitive Processes—Knowledge Sources and Blackboard System

The cognitive components of the system may be implemented using a *blackboard architecture*—for details see Schymura and Kolossa (2020), this volume, and Brown et al. (2014). The blackboard architecture includes expert modules, so-called *knowledge sources* (KSs). These carry out specific analysis tasks, such as lower-level pre-segmentation, source separation, visual-pattern detection and tracking, that is, the involved knowledge sources act in terms of low-level experts for pre-segmentation and *Gestalt*-type analysis. Higher-level KSs as experts for tasks such as detecting, classifying and labeling sound events. At a higher level, knowledge sources need to be implemented that assign meaning to perceptual objects and to the auditory events they are associated with. The methods of each level pass their output information on to the blackboard system. Higher-layer experts use this information and related statistical uncertainty data to generate hypotheses. At the very highest layer, cognitive processes need to be implemented, whereby their expertise includes world knowledge (Brown et al. 2014).

At the intersection between blackboard events and knowledge sources, the focusing of attention takes place (Brown et al. 2014; Schymura and Kolossa 2020), this volume. This may comprise the selection of specific blackboard information by KSs, or of specific types of input information from the sensory representation. It may also involve top-down feedback, for example, adjusting the filter bandwidths of the basilar membrane to a specific kind of input signal or triggering head-motion to direct the head to a certain scene object. Across all layers, the expertise provided by the different experts includes, among other fields of knowledge, psychoacoustics, object-identification, cross-modal integration, proprioception with regard to head- and general movements, speech communication-specific expertise such as speech-versus noise-identification and word recognition, music identification and classification, and sound quality evaluation.

¹²See <http://docs.twoears.eu/en/1.5/afe/> [last accessed: February 22, 2020].

Feedback Mechanisms

In human audition, as part of human perception and cognition, feedback serves to improve certain performances, such as object recognition, auditory grouping, aural-stream segregation, scene analysis, and hence improve the scene understanding, assignment of meaning, attention focusing, and also the evaluation of *Sound Quality* and *Quality of Experience*. Feedback mechanisms involve both a process that is initiating feedback information and another process that receives and acts upon it—for details refer to Blauert and Brown (2020), this volume.

5.3 Benefits of Holistic Hearing Model for Sound Quality and Quality of Experience Models

Applied to *Sound Quality* and *Quality of Experience* estimation, such models may provide the following functional capabilities (Raake and Blauert 2013; Raake et al. 2014b).

Learned internal references rather than explicit reference signals. With a corresponding *no-reference* sound quality model, the quality can be directly estimated based on the available ear-signals. Moreover, also for a model that uses a reference signal—that is, a so-called *full-reference* model—a functionally adequate reference-adaptation may be addressed. Two different approaches are conceivable, that is, (i) rule-based approaches with a restricted dataset available for model and reference training—for example combining multidimensional analysis with a preference-mapping-type relation to *Sound Quality* or QoE—see Sects. 3.2, 4—and, (ii) data-based approaches, where some kind of learning of references is involved or transfer learning is applied—see Spille et al. (2018) and Göring et al. (2018). Larger datasets may be established for a direct training of Deep-Neural-Network-(DNN)-type models instead of transfer-learning using, for example, quality ratings as they are collected, e.g., by Skype in the field after selected calls, or via crowd-sourcing¹³ (Hossfeld et al. 2014).

Identification of scene and source types and respective adjustment of low-level processing as well as adjustment of the selected internal reference, in light of the given evaluation task and acoustic scene. For example, music or speech may be recognized as the primary input. Appropriate pre-trained machine-learning models may then be used for genre recognition or speech intelligibility estimation.

Scene-object-specific evaluation with multiple objects being present in an auditory scene. Quality evaluation will then be scene- and object-specific (e.g., see, Raake et al. 2014b). Such a scene-based quality-modeling paradigm is principally enabled by a model that includes a dedicated scene-analysis stage. Some

¹³Crowd-sourcing tests involving dedicated crowd-workers are distinguished from data collection in the field with a more arbitrary and hence real-life sample of users, and with a less guided, more natural usage behavior.

first considerations along these lines for sound quality using scene foreground- and background-related features have been proposed in Skowronek et al. (2017). *Implementation of attentional processes* based on the scene- and object-oriented paradigm. In this way, saliency and selective attention can be incorporated into the model. First approaches along these lines for the existing TWO!EARS framework are described in Cohen l’Hyver (2017), and Cohen-L’Hyver et al. (2020), this volume, but have not yet been applied to *Sound Quality* and *Quality of Experience* modeling. An attention model for soundscapes has been presented in Oldoni et al. (2013).

Integration with visual information, in terms of specific features of the scene (Cohen l’Hyver 2017). In this way, the adaptation of lower-level processing as, for example, related to the precedence effect, may be included (Braasch 2020, this volume). Further, aspects such as the visual and auditory congruency of the room and the respective role for externalization may be addressed, an effect referred to as *room divergence* (Werner et al. 2016; Brandenburg et al. 2020, this volume).

Active exploration enabling the model to explore the auditory scene and include the exploration for an improved or simply more human-like assessment such as, (i) targeting a specific analysis of certain low-level features exploited during interactive quality evaluation, for example, based on behavioral patterns, or (ii) enabling the exploration of the scene, for example, to identify the sweet-spot of a given sound reproduction system in a perceptual way. This is complementary to the experimental work described in Kim et al. (2013) and informal experiments performed by the authors during the TWO!EARS project (cf. www.twoears.eu).

With such an underlying active listening model, *Sound Quality* and *Quality of Experience* modeling can be based on a running sound quality-feature model, using a combination of a set of cue-analysis components. Higher model layers could include quality-feature integration, and additional high-level components that are able to generate top-down events that includes other factors, such as the liking/disliking of a given piece of music, the focus of attention of the listener, or the visual information provided in addition to the auditory information.

It is clear that at this stage, such a model does not exist, and work reported so far only implements parts of these concepts (e.g., Raake and Wierstorf 2016; Skowronek et al. 2017).

6 Conclusions and Future Directions

The current chapter discussed different concepts related to *Sound Quality* and the more holistic, yet harder to assess, *Quality of Experience*. Respective assessment methods were summarized in light of these concepts. Based on the work conducted in the TWO!EARS project, a conceptual *Sound Quality* and *Quality of Experience* model was introduced. The model components were outlined, and it was analyzed how different types of quality-related models can be implemented with these. Previously,

it has been shown that this approach enables the design of quality-feature models for coloration and localization prediction (Raake et al. 2014b; Raake and Wierstorf 2016) as well as for preference prediction (Skowronek et al. 2017).

Open-source availability of algorithms and data is one of the key challenges for audio-quality research and modeling. Most well-established existing model approaches such as POLQA (ITU–T Rec. P.863 2011), QESTRAL (Rumsey et al. 2008), or PEMO-Q (Harlander et al. 2014) are proprietary, and no explicit source code has been made available. Some few attempts for reverse-engineering exist, for example, with the PEASS *Perceptual Evaluation methods for Audio Source Separation* toolkit (Emiya et al. 2011) or via the code in the Github project *Perceptual coding in Python*.¹⁴ The open-source *Auditory Front End* (AFE) of TWO!EARS¹⁵ was developed by applying elements from the open-source *Auditory Modeling Toolbox* (AMT).¹⁶

An approach for an explicit collaborative model development could be enabled by reproducible research around toolboxes such as the AMT that are worked on by a larger community. Here, it will be helpful if public funding agencies foster activities that emphasize such fundamental though practical inter-group collaborations. Further, it should be more widely accepted in the scientific community that “toolboxes” actually represent (even highly valuable) scientific work, too.

Auditory perception research—as part of *Sound Quality* and *Quality of Experience* evaluation—could certainly be advanced at large with the help of high-quality toolboxes. Yet, to be sure, such endeavor must be based on a deep understanding of auditory perception and requires profound software-development skills. The final goal is to achieve a well documented, tested and ultimately widely adopted basis for future scientific discoveries.

As was highlighted by an analysis of recent tests on *Sound Quality* and *Quality of Experience*, in Sect. 3.5, very few databases are publicly available that could be used for model training. Of course, the creation and sharing of databases could go hand in hand with a collaborative model development project as it was advocated above. To this aim, already the sharing of known proprietary databases (e.g., see the list in Table 3) would be a very welcome contribution to the domain of perceptual sound quality and QoE modeling.

In the current chapter, it was discussed how actual model implementations can be trained with listening-test data. Here, different approaches, especially for the training of internal model knowledge and internal references, were considered. Limitations were highlighted that currently reduce the feasibility of developing a full *Sound Quality* and *Quality of Experience* model.

Besides the challenges involved when developing a basic-quality model, the question arises of how the different *contexts* as discussed in Sect. 2.5 and, hence, also

¹⁴<https://github.com/stephencwelch/Perceptual-Coding-In-Python> [last accessed: August 30, 2019].

¹⁵ <http://docs.twoears.eu/en/1.5/afe/> [last accessed: August 30, 2019].

¹⁶ Søndergaard et al. (2011) and Søndergaard and Majdak (2013), <http://amtoolbox.sourceforge.net/> [last accessed: August 31, 2019].

individual differences can be implemented in a perception model. This aspect is a highly relevant issue to be solved since the context-specific evaluation of audio and especially spatial audio is an important requirement for ecological validity. For example, features such as envelopment (e.g., compare Rumsey 2002) will be differently desirable depending on the given context. In the current authors' opinion, this aspect is one of the biggest challenges in *Sound Quality* and *Quality of Experience* modeling.

Reflecting listener-internal references and a system/scene-control as discussed in Sect. 2.6 in a quality model, and this in a person- and expertise-specific manner, appears to be still out of reach. Nevertheless, it represents a rewarding goal for a better understanding of human perception and evaluation as well as for the application of the resulting models for automatic audio-system adaptation and optimization.

Acknowledgements This research has partly been supported by EU-FET grant TWO!EARS, ICT-618075. The authors are grateful to Chris Hold, Marie-Neige Garcia, Werner Robitza, Sebastian Egger, Sebastian Möller, John Mourjopoulos, Sascha Spors, Karlheinz Brandenburg, Janina Fels, and Patrick Danès for fruitful discussions and conceptual contributions. Two external reviewer have provided useful comments and advice for improving this chapter.

References

- Amelang, M., D.G.S. Bartussek, and D. Hagemann. 2006. *Differentielle Psychologie und Persönlichkeitsforschung (Differential Psychology and Personality Research)*. Stuttgart: W. Kohlhammer Verlag.
- Baddeley, A. 1997. *Human Memory—Theory and Practice*. East Sussex, UK: Taylor & Francis, Psychology Press.
- Baddeley, A. 2003. Working memory: Looking back and looking forward. *Nature Reviews Neuroscience* 4: 829–839. <https://doi.org/10.1038/nrn1201>.
- Bech, S., and N. Zacharov. 2006. *Perceptual Audio Evaluation*. Chichester, UK: Wiley.
- Beerends, J.G., C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl. 2013. Perceptual Objective Listening Quality Assessment (POLQA), The third generation ITU-T standard for end-to-end speech quality measurement. Part II—Perceptual model. *Journal of the Audio Engineering Society* 61 (6): 385–402. <http://www.aes.org/e-lib/browse.cfm?elib=16829>. Accessed 9 Oct 2019.
- Benoit, A., P. LeCallet, P. Campisi, and R. Cousseau. 2008. Quality assessment of stereoscopic images. In *IEEE International Conference Image Processing (ICIP)* 1231–1234.
- Bentham, J. 1789. *An Introduction to the Principle of Morals and Legislations*. Oxford, UK: Blackwell (Reprint 1948).
- Biberger, T., and S.D. Ewert. 2016. Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility. *The Journal of the Acoustical Society of America* 140 (2): 1023–1038. <https://doi.org/10.1121/1.4960574>.
- Biberger, T., J.-H. Fleßner, R. Huber, and S.D. Ewert. 2018. An objective audio quality measure based on power and envelope power cues. *Journal of the Audio Engineering Society* 66 (7/8), 578–593. <http://www.aes.org/e-lib/browse.cfm?elib=19707>. Accessed 23 Sept 2019.
- Blauert, J. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: The MIT Press.
- Blauert, J. 2013. Conceptual aspects regarding the qualification of spaces for aural performances. *Acta Acustica united with Acustica* 99: 1–13. <https://doi.org/10.3813/AAA.918582>.

- Blauert, J. 2020. A virtual testbed for binaural agents. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 491–510. Cham, Switzerland: Springer and ASA Press.
- Blauert, J., J. Braasch, J. Buchholz, H.S. Colburn, U. Jekosch, A. Kohlrausch, J. Mourjopoulos, V. Pulkki, and A. Raake. 2009. Aural assessment by means of binaural algorithms – the AABBA project. In *Proceedings of the 2nd International Symposium Auditory and Audiological Research–ISAAR’09*, 113–124.
- Blauert, J., and G. Brown. 2020. Reflexive and reflective auditory feedback. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 3–31, Cham, Switzerland: Springer and ASA Press. This volume.
- Blauert, J., and U. Jekosch. 2012. A layer model of sound quality. *Journal of the Audio Engineering Society* 60 (1/2): 4–12. <http://www.aes.org/e-lib/browse.cfm?elib=16160>. Accessed 19 Sept 2019.
- Blauert, J., D. Kolossa, K. Obermayer, and K. Adiloglu. 2013. Further challenges—and the road ahead. In *The Technology of Binaural Listening*, ed. J. Blauert. Berlin: Springer and ASA Press. https://doi.org/10.1007/978-3-642-37762-4_18.
- Bowman, D.A., and R.P. McMahan. 2007. Virtual reality: How much immersion is enough? *Computer* 40 (7): 36–43.
- Braasch, J. 2020. Binaural modeling from an evolving habitat perspective. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 251–286, Cham, Switzerland: Springer and ASA Press.
- Bradley, J.S., and G.A. Soulodre. 1995. Objective measures of listener envelopment. *Journal of the Acoustical Society of America* 98 (5): 2590–2597.
- Brandenburg, K., F. Klein, A. Neidhardt, U. Sloma, and S. Werner. 2020. Creating auditory illusions with binaural technology. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 623–663, Cham, Switzerland: Springer and ASA Press.
- Bregman, A.S. 1990. *Auditory Scene Analysis*. Cambridge, USA: The MIT Press.
- Brown, G., R. Decorsière, D. Kolossa, N. Ma, T. May, C. Schymura, and I. Trowitzsch. 2014. *D3.1: TWO!EARS Software Architecture, Two!Ears FET-Open Project*. <https://doi.org/10.5281/zenodo.2595254>.
- Brüggen, M. 2001a. Coloration and binaural decoloration in natural environments. *Acta Acustica united with Acustica* 87: 400–406.
- Brüggen, M. 2001b. Sound coloration due to reflections and its auditory and instrumental compensation. PhD thesis, Ruhr-Universität Bochum.
- Carroll, J.D. 1972. Individual preferences and multidimensional scaling. In *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences*, vol. I, ed. R.N. Shepard, A.K. Romney, and S.B. Nerlove, 105–155.
- Choisel, S., and F. Wickelmaier. 2007. Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. *The Journal of the Acoustical Society of America* 121 (1): 388–400. <https://doi.org/10.1121/1.2385043>.
- Cohen l’Hyver, B. 2017. Modulation de mouvements de tête pour l’analyse multimodale d’un environnement inconnu (modulation of head movements for the multimodal analysis of an unknown environment). PhD thesis, Université Pierre et Marie Curie, Ecole Doctorale SMAER, Sciences Mécaniques, Acoustique, Electronique et Robotique de Paris, France.
- Cohen-L’Hyver, B., S. Argentieri, and B. Gas. 2020. Audition as a trigger of head movements. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 697–731, Cham, Switzerland: Springer and ASA Press.
- Coltheart, M. 1980. Iconic memory and visible persistence. *Perception & Psychophysics* 27 (3): 183–228. <https://doi.org/10.3758/BF03204258>.
- Cowan, N. 1984. On short and long auditory stores. *Psychol. Bulletin* 96 (2): 341–370. <https://doi.org/10.1037/0033-2909.96.2.341>.
- Dobrian, F., A. Awan, D. Joseph, A. Ganjam, J. Zhan, V. Sekar, I. Stioca, and H. Zhang. 2013. Understanding the impact of video quality on user engagement. *Communications of the ACM* 56 (3): 91–99. <https://doi.org/10.1145/2043164.2018478>.

- Emiya, V., E. Vincent, N. Harlander, and V. Hohmann. 2011. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing* 19 (7): 2046–2057. <https://doi.org/10.1109/TASL.2011.2109381>.
- Engelke, U., D.P. Darcy, G.H. Mulliken, S. Bosse, M.G. Martini, S. Arndt, J.-N. Antons, K.Y. Chan, N. Ramzan, and K. Brunnström. 2017. Psychophysiology-based qoe assessment: A survey. *IEEE Journal of Selected Topics in Signal Processing* 11 (1): 6–21. <https://doi.org/10.1109/JSTSP.2016.2609843>.
- Francombe, J., T. Brookes, and R. Mason. 2017a. Evaluation of spatial audio reproduction methods (part 1): Elicitation of perceptual differences. *Journal of the Audio Engineering Society* 65 (3): 198–211. <https://doi.org/10.17743/jaes.2016.0070>.
- Francombe, J., T. Brookes, R. Mason, and J. Woodcock. 2017b. Evaluation of spatial audio reproduction methods (part 2): Analysis of listener preference. *Journal of the Audio Engineering Society* 65 (3): 212–225. <https://doi.org/10.17743/jaes.2016.0071>.
- Garcia, M.-N., R. Schleicher, and A. Raake. 2011. Impairment-factor-based audiovisual quality model for iptv: Influence of video resolution, degradation type, and content type. *EURASIP Journal on Image and Video Processing* 2011 (1): 1–14. <https://doi.org/10.1155/2011/629284>.
- Geerts, D., K.D. Moor, I. Ketyko, A. Jacobs, J.V. den Bergh, W. Joseph, L. Martens, and L.D. Marez. 2010. Linking an integrated framework with appropriate methods for measuring QoE. In *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEX)*. <https://doi.org/10.1109/QOMEX.2010.5516292>.
- Görling, S., J. Skowronek, and A. Raake. 2018. DeViQ - A deep no reference video quality model. In *Proceedings Human Vision and Electronic Imaging (HVEI)* 1–6: <https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-518>.
- Griesinger, D. 1998. General overview of spatial impression, envelopment, localization, and externalization. In *Audio Engineering Society Conference: 15th International Conference: Audio, Acoustics & Small Spaces*, *Audio Engineering Society*. <http://www.aes.org/e-lib/browse.cfm?elib=8095>. Accessed 17 Sept 2019.
- Harlander, N., R. Huber, and S.D. Ewert. 2014. Sound quality assessment using auditory models. *Journal of the Audio Engineering Society* 62 (5): 324–336. <https://doi.org/10.17743/jaes.2014.0020>.
- Härmä, A., M. Park, and A. Kohlrausch. 2014. Data-driven modeling of the spatial sound experience. In *Audio Engineering Society Convention 136*. <http://www.aes.org/e-lib/browse.cfm?elib=17172>. Accessed 18 Sept 2019.
- Hartmann, W.M., and A. Wittenberg. 1996. On the externalization of sound images. *Journal of the Acoustical Society of America* 99 (6): 3678–3688.
- Hassenzahl, M. 2001. The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction* 13 (4): 481–499. https://doi.org/10.1207/S15327590IJHC1304_07.
- Hossfeld, T., C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia. 2014. Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. *IEEE Transactions on Multimedia* 16 (2): 541–558. <https://doi.org/10.1109/TMM.2013.2291663>.
- Houtgast, T., and H.J.M. Steeneken. 1985. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America* 77 (3): 1069–1077. <https://doi.org/10.1121/1.392224>.
- ISO 9000:2000. 2000. *Quality Management Systems: Fundamentals and Vocabulary*, International Organization for Standardization.
- ITU-R BS. 1116-1. 1997. *Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems*. Geneva, CH: International Telecommunication Union.
- ITU-R BS. 1283-1. 2003. *A Guide to ITU-R Recommendations for Subjective Assessment of Sound Quality*. Geneva, CH: International Telecommunication Union.
- ITU-R BS. 1534-3. 2015. *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*. Geneva, CH: International Telecommunication Union.

- ITU-T Rec. P.10/G.100. 2017. *Vocabulary for Performance and Quality of Service*. Geneva, CH: International Telecommunication Union.
- ITU-T Rec. P.800. 1996. *Methods for Subjective Determination of Transmission Quality*. Geneva, CH: International Telecommunication Union.
- ITU-T Rec. P.862. 2001. *Perceptual Evaluation of Speech Quality (PESQ)*, International Telecommunication Union.
- ITU-T Rec. P.863. 2011. *Perceptual Objective Listening Quality Assessment (POLQA)*, International Telecommunication Union.
- Jekosch, U. 2004. Basic concepts and terms of “quality”, reconsidered in the context of product sound quality. *Acta Acustica united with Acustica* 90 (6): 999–1006.
- Jekosch, U. 2005a. Assigning meaning to sounds: Semiotics in the context of product-sound design. In *Communication Acoustics*, ed. J. Blauert. Berlin: Springer. https://doi.org/10.1007/3-540-27437-5_8.
- Jekosch, U. 2005b. *Voice and Speech Quality Perception—Assessment and Evaluation*. D-Berlin: Springer.
- Kahneman, D. 1999. Objective happiness. In *Well-Being: The Foundations of Hedonic Psychology*, ed. D. Kahneman, E. Diener, and N. Schwarz, 3–25. New York: Russell Sage Foundation.
- Kahneman, D. 2003. Experienced utility and objective happiness: A moment-based approach. In *The Psychology of Economic Decisions*, ed. I. Brocas, and J.D. Carrillo, 187–208. Oxford: Oxford University Press.
- Kahneman, D. 2011. *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.
- Kim, C., R. Mason, and T. Brookes. 2013. Head movements made by listeners in experimental and real-life listening activities. *Journal of the Audio Engineering Society* 61 (6): 425–438. <http://www.aes.org/e-lib/browse.cfm?elib=16833>. Accessed 18 Sept 2019.
- Kuttruff, H. 2016. *Room Acoustics*. Boca Raton: CRC Press.
- Lawless, H.T., and H. Heymann. 2010. *Sensory Evaluation of Food: Principles and Practices*, vol. 5999. Berlin: Springer.
- Lebreton, P., A. Raake, M. Barkowsky, and P.L. Callet. 2013. Perceptual preference of S3D over 2D for HDTV in dependence of video quality and depth. In *IVMSP Workshop: 3D Image/Video Technologies and Applications, 10–12 June*, 1–4. Korea, Seoul.
- Lepa, S., E. Ungeheuer, H.-J. Maempel, and S. Weinzierl. 2013. When the medium is the message: An experimental exploration of medium effects on the emotional expressivity of music dating from different forms of spatialization. In *Proceedings of the 8th Conference of the Media Psychology Division of Deutsche Gesellschaft für Psychologie (DGPs)*.
- Lepa, S., S. Weinzierl, H.-J. Maempel, and E. Ungeheuer. 2014. Emotional impact of different forms of spatialization in everyday mediated music listening: Placebo or technology effects? In *Audio Engineering Society Convention 136*, Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=17171>. Accessed 18 Sept 2019.
- Letowski, T. 1989. Sound quality assessment: Concepts and criteria. In *Audio Engineering Society Convention 87, 18–21 Oct*, New York, USA. <http://www.aes.org/e-lib/browse.cfm?elib=5869>. Accessed 18 Sept 2019.
- Li, J., M. Barkowsky, and P. LeCallet. 2012. Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment. In *IEEE International Conference Image Processing (ICIP), 30 Sept–03 Oct*, Orlando, Florida, USA.
- Liebetrau, J., T. Sporer, S. Kämpf, and S. Schneider. 2010. Standardization of PEAQ-MC: Extension of ITU-R BS.1387-1 to multichannel audio. In *Audio Engineering Society, 40th International Conference: Spatial Audio, 8–10 Oct*, Tokyo, Japan. <http://www.aes.org/e-lib/browse.cfm?elib=15571>. Accessed 23 Sept 2019.
- Lindau, A., V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl. 2014. A spatial audio quality inventory (SAQI). *Acta Acustica united with Acustica* 100 (5): 984–994. <https://doi.org/10.3813/AAA.918778>.

- Lokki, T., J. Pätynen, A. Kuusinen, H. Vertanen, and S. Tervo. 2011. Concert hall acoustics assessment with individually elicited attributes. *The Journal of the Acoustical Society of America* 130 (2): 835–849. <https://doi.org/10.1121/1.3607422>.
- Martens, H., and M. Martens. 2001. *Multivariate Analysis of Quality*. Chichester: Wiley.
- Massaro, D.W. 1975. Backward recognition masking. *The Journal of the Acoustical Society of America* 58 (5): 1059–1065. <https://doi.org/10.1121/1.380765>.
- Mattila, V. 2001. *Perceptual Analysis of Speech Quality in Mobile Communications*, vol. 340. Doctoral Dissertation, Tampere University of Technology, FIN–Tampere.
- Mattila, V. 2002. Ideal point modelling of speech quality in mobile communications based on multidimensional scaling. Audio Engineering Society Convention, vol. 112. <http://www.aes.org/e-lib/browse.cfm?elib=11433>. Accessed 23 Sept 2019.
- Mausfeld, R. 2003. Conjoint representations and the mental capacity for multiple simultaneous perspectives. In *Looking into Pictures: An Interdisciplinary Approach to Pictorial Space*, ed. H. Hecht, R. Schwartz, and M. Atherton, 17–60. Cambridge: MIT Press.
- Moor, K.D. 2012. Are engineers from mars and users from venus? Bridging the gaps in quality of experience research: Reflections on and experiences from an interdisciplinary journey. PhD thesis, Universiteit Gent.
- Moore, B.C., and B.R. Glasberg. 2007. Modeling binaural loudness. *The Journal of the Acoustical Society of America* 121 (3): 1604–1612. <https://doi.org/10.1121/1.2431331>.
- Moore, B.C.J., and C.-T. Tan. 2004. Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion. *Journal of the Audio Engineering Society* 52 (9): 900–914. <http://www.aes.org/e-lib/browse.cfm?elib=13018>. Accessed 23 Sept 2019.
- Mourjopoulos, J. 2020. Aesthetics aspects regarding recorded binaural sounds. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 455–490, Cham, Switzerland: Springer and ASA Press.
- Neisser, U. 1978. Perceiving, anticipating and imagining. *Minnesota Studies in the Philosophy of Science* 9: 89–106.
- Neisser, U. 1994. Multiple systems: A new approach to cognitive theory. *European Journal of Cognitive Psychology* 6 (3): 225–241. <https://doi.org/10.1080/09541449408520146>.
- Oldoni, D., B. De Coensel, M. Boes, M. Rademaker, B. De Baets, T. Van Renterghem, and D. Botteldooren. 2013. A computational model of auditory attention for use in soundscape research. *The Journal of the Acoustical Society of America* 134 (1): 852–861. <https://doi.org/10.1121/1.4807798>.
- Olko, M., D. Dembeck, Y.-H. Wu, A. Genovese, and A. Roginska. 2017. Identification of perceived sound quality attributes of 360-degree audiovisual recordings in VR – Using a free verbalization method. In *Audio Engineering Society Convention 143, 18–21 Oct*, New York, USA. Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=19227>. Accessed 23 Sept 2019.
- Parasuraman, A., V. Zeithaml, and L. Berry. 1985. A conceptual model of service quality and its implications for future research. *Journal of Marketing* 49 (Fall 1985): 41–50. <https://doi.org/10.2307/1251430>.
- Piaget, J. 1962. *The Child's Conception of the World (La représentation du monde chez l'enfant)*. London: Routledge & Kegan. Translated from the 1926 original.
- Pike, C., and H. Stenzel. 2017. Direct and indirect listening test methods – A discussion based on audio-visual spatial coherence experiments. In *Audio Engineering Society Convention 143*, Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=19226>. Accessed 23 Sept 2019.
- QUALINET. 2012. *White Paper on Definitions of Quality of Experience*, COST Action IC 1003, ed. Möller, S., P. Le Callet, and A. Perkis, Lausanne, CH
- Raake, A. 2006. *Speech Quality of VoIP–Assessment and Prediction*. Chichester, West Sussex, UK: Wiley.
- Raake, A. 2016. Views on sound quality. In *Proceedings 22nd International Congress on Acoustics (ICA), 5–9 Sept*, 1–10, Buenos Aires, Argentina.

- Raake, A., and J. Blauert. 2013. Comprehensive modeling of the formation process of sound-quality. In *Proceedings of the IEEE International Conference Quality of Multimedia Experience (QoMEX)*, 3–5 July, Klagenfurt, Austria. <https://doi.org/10.1109/QoMEX.2013.6603214>.
- Raake, A., J. Blauert, J. Braasch, G. Brown, P. Danes, T. Dau, B. Gas, S. Argentieri, A. Kohlrausch, D. Kolossa, N. Le Goeff, T. May, K. Obermayer, C. Schymura, T. Walther, H. Wierstorf, F. Winter, and S. Spors. 2014a. Two!ears – Integral interactive model of auditory perception and experience. In *40th German Annual Conference on Acoustics (DAGA)*, 10–13 March, Oldenburg, Germany.
- Raake, A., H. Wierstorf, and J. Blauert. 2014b. A case for Two!Ears in audio quality assessment. *Forum Acusticum*, 7–12 Sept., Krakow, Poland.
- Raake, A., and S. Egger. 2014. Quality and quality of experience. In *Quality of Experience. Advanced Concepts, Applications and Methods*, ed. S. Möller, and A. Raake. Berlin: Springer. Chap. 2. https://doi.org/10.1007/978-3-319-02681-7_2.
- Raake, A., C. Schlegel, K. Hoeldtke, M. Geier, and J. Ahrens. 2010. Listening and conversational quality of spatial audio conferencing. In *40th International Conference on Spatial Audio: Sense the Sound of Space*, Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=15567>. Accessed 23 Sept 2019.
- Raake, A., and H. Wierstorf. 2016. Assessment of audio quality and experience using binaural-hearing models. In *Proceedings 22nd International Congress on Acoustics (ICA)*, 5–9 Sept., 1–10. Buenos Aires, Argentina.
- Reardon, G., A. Genovese, G. Zalles, P. Flanagan, and A. Roginska. 2018. Evaluation of binaural renderers: Multidimensional sound quality assessment. In *2018 International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=19694>. Accessed 23 Sept 2019.
- Rees-Jones, J., and D.T. Murphy. 2018. The impact of multichannel game audio on the quality and enjoyment of player experience. In *Emotion in Video Game Soundtracking*, 143–163. Berlin: Springer. https://doi.org/10.1007/978-3-319-72272-6_11.
- Reeves, C.A., and D.A. Bednar. 1994. Defining quality: Alternatives and implications. *Academy of Management Review* 19 (3): 419–445. <https://doi.org/10.2307/258934>.
- Reiter, U., K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank. 2014. Factors influencing quality of experience. In *Quality of Experience. Advanced Concepts, Applications and Methods*, ed. S. Möller, and A. Raake. Berlin: Springer. Chap. 4. https://doi.org/10.1007/978-3-319-02681-7_4.
- Richards, D.L. 1973. *Telecommunication by Speech*. London, UK: Butterworths.
- Richards, D.L. 1973. *Telecommunication by Speech*. London, UK: Butterworths.
- Rummukainen, O., T. Robotham, S.J. Schlecht, A. Plinge, J. Herre, and E.A. Habets. 2018. Audio quality evaluation in virtual reality: Multiple stimulus ranking with behavior tracking. In *2018 AES International Conference on Audio for Virtual and Augmented Reality*, Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=19678>. Accessed 23 Sept 2019.
- Rummukainen, O., S. Schlecht, A. Plinge, and E.A. Habets. 2017. Evaluation of binaural reproduction systems from behavioral patterns in a six-degrees-of-freedom wayfinding task. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 1–3. <https://doi.org/10.1109/QoMEX.2017.7965680>.
- Rumsey, F. 2002. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society* 50 (9): 651–666. <http://www.aes.org/e-lib/browse.cfm?elib=11067>. Accessed 23 Sept 2019.
- Rumsey, F., S. Zieliński, P. Jackson, M. Dewhirst, R. Conetta, S. George, S. Bech, and D. Mearns. 2008. QESTRAL (part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener. In *Audio Engineering Society Convention 125*, 3–5 Oct, San Francisco, USA. <http://www.aes.org/e-lib/browse.cfm?elib=14746>. Accessed 23 Sept 2019.
- Rumsey, F., S. Zieliński, R. Kassier, and S. Bech. 2005. On the relative importance of spatial and timbral fidelities in judgements of degraded multichannel audio quality. *Journal of the Acoustical Society of America* 118 (2): 968–976. <https://doi.org/10.1121/1.1945368>.

- Schoeffler, M., and J. Herre. 2013. About the impact of audio quality on overall listening experience. In *Proceedings of the Sound and Music Computing Conference (SMC)*, 30 July–3 Aug., Stockholm, Sweden, 53–58.
- Schoeffler, M., and J. Herre. 2016. The relationship between basic audio quality and overall listening experience. *Journal of the Acoustical Society of America* 140 (3): 2101–2112. <https://doi.org/10.1121/1.4963078>.
- Schoeffler, M., A. Silzle, and J. Herre. 2017. Evaluation of spatial/3d audio: Basic audio quality versus quality of experience. *IEEE Journal of Selected Topics in Signal Processing* 11 (1): 75–88. <https://doi.org/10.1109/JSTSP.2016.2639325>.
- Schoenenberg, K. 2016. The quality of mediated-conversations under transmission delay. PhD thesis, Technische Universität Berlin. <https://doi.org/10.14279/depositonce-4990>.
- Schoenenberg, K., A. Raake, and J. Koeppel. 2014. Why are you so slow?—misattribution of transmission delay to attributes of the conversation partner at the far-end. *International Journal of Human-Computer Studies* 72 (5): 477–487. <https://doi.org/10.1016/j.ijhcs.2014.02.004>.
- Schymura, C., and D. Kolossa. 2020. Blackboard systems for cognitive audition. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 91–111, Cham, Switzerland: Springer and ASA Press. Chap. 4.
- Seo, J.-H., S.B. Chon, K.-M. Sung, and I. Choi. 2013. Perceptual objective quality evaluation method for high-quality multichannel audio codecs. *Journal of the Audio Engineering Society* 61 (7/8): 535–545. <http://www.aes.org/e-lib/browse.cfm?elib=16869>. Accessed 23 Sept 2019.
- Singla, A., S. Fremerey, W. Robitza, and A. Raake. 2017. Measuring and comparing goe and simulator sickness of omnidirectional videos in different head mounted displays. In *Proceedings of the International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, Germany. IEEE, 1–6. <https://doi.org/10.1109/QoMEX.2017.7965658>.
- Skowronek, J., L. Nagel, C. Hold, H. Wierstorf, and A. Raake. 2017. Towards the development of preference models accounting for the impact of music production techniques. In *43rd German Annual Conference on Acoustics (DAGA)*, 856–860.
- Skowronek, J., and A. Raake. 2015. Assessment of cognitive load, speech communication quality and quality of experience for spatial and non-spatial audio conferencing calls. *Speech Communication* 66: 154–175. <https://doi.org/10.1016/j.specom.2014.10.003>.
- Søndergaard, P., J. Culling, T. Dau, N. Le Goff, M. Jepsen, P. Majdak, and H. Wierstorf. 2011. Towards a binaural modelling toolbox. In *Proceedings of the Forum Acusticum, European Acoustics Association (EAA)*, 27 June–01 July, Aalborg, Denmark, 2081–2086.
- Søndergaard, P., and P. Majdak. 2013. The auditory-modeling toolbox. In *The Technology of Binaural Listening*, ed. J. Blauert. Berlin: Springer and ASA Press. Chap. 2. https://doi.org/10.1007/978-3-642-37762-4_2.
- Sotujo, S., J. Thiemann, A. Kohlrausch, and S. Van de Paar. 2020. Auditory gestalt rules and their application. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 33–59, Cham, Switzerland: Springer and ASA Press.
- Spille, C., S.D. Ewert, B. Kollmeier, and B. Meyer. 2018. Predicting speech intelligibility with deep neural networks. *Computer Speech & Language* 48: 51–66. <https://doi.org/10.1016/j.csl.2017.10.004>.
- Spors, S., M. Geier, and H. Wierstorf. 2017. Towards open science in acoustics: Foundations and best practices. In *Proceedings of the 43. Jahrestagung f. Akustik (43th Annual Meeting German Society Acoustics, DAGA)*, 6–9 March, Kiel, Germany, 218–221.
- Spors, S., H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter. 2013. Spatial sound with loudspeakers and its perception: A review of the current state. *Proceedings of the IEEE* 101 (9): 1920–1938. <https://doi.org/10.1109/JPROC.2013.2264784>.
- Strohmeier, D., S. Jumisko-Pyykkö, and K. Kunze. 2010. Open profiling of quality: A mixed method approach to understanding multimodal quality perception. *Advances in Multimedia* 2010 (Article ID 658980): 28. <https://doi.org/10.1155/2010/658980>.
- Thiede, T., W. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten. 2000. PEAQ—the ITU standard for objective measurement

- of perceived audio quality. *Journal of the Audio Engineering Society* 48: 3–29. <http://www.aes.org/e-lib/browse.cfm?elib=12078>. Accessed 23 Sept. 2019.
- Uhrig, S., S. Arndt, S. Möller, and J. Voigt-Antons. 2017. Perceptual references for independent dimensions of speech quality as measured by electro-encephalography. *Quality and User Experience* 2 (1): 1–10. <https://doi.org/10.1007/s41233-017-0011-8>.
- Uhrig, S., G. Mittag, S. Möller, and J.-N. Voigt-Antons. 2018. Neural correlates of speech quality dimensions analyzed using electroencephalography (EEG). *Journal of Neural Engineering*.
- van Ee, R., J.J.A. van Boxtel, A.L. Parker, and D. Alais. 2009. Multisensory congruency as a mechanism for attentional control over perceptual selection. *Journal of Neuroscience* 29 (37): 11641–11649. <https://doi.org/10.1523/JNEUROSCI.0873-09.2009>.
- Wältermann, M. 2005. Bestimmung relevanter Qualitätsdimensionen bei der Sprachübertragung in modernen Telekommunikationsnetzen. Diploma thesis (unpublished), Institut für Kommunikationsakustik, Ruhr-Universität, D-Bochum.
- Wältermann, M. 2013. *Dimension-Based Quality Modeling of Transmitted Speech*. Berlin: Springer Science & Business Media.
- Wältermann, M., A. Raake, and S. Möller. 2010. Quality dimensions of narrowband and wideband speech transmission. *Acta Acustica united with Acustica* 96 (6): 1090–1103. <https://doi.org/10.3813/AAA.918370>.
- Weiss, B., D. Guse, S. Möller, A. Raake, A. Borowiak, and U. Reiter. 2014. Temporal development of quality of experience. In *Quality of Experience. Advanced Concepts, Applications and Methods*, ed. S. Möller, and A. Raake, 133–147. Berlin: Springer. Chap. 10. https://doi.org/10.1007/978-3-319-02681-7_10.
- Werner, S., F. Klein, T. Mayenfels, and K. Brandenburg. 2016. A summary on acoustic room divergence and its effect on externalization of auditory events. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 1–6. <https://doi.org/10.1109/QoMEX.2016.7498973>.
- Wickelmaier, F., N. Umbach, K. Sering, and S. Choisel. 2009. Comparing three methods for sound quality evaluation with respect to speed and accuracy. In *Audio Engineering Society Convention 126*, Audio Engineering Society.
- Wierstorf, H. 2014. Perceptual assessment of sound field synthesis. PhD thesis, TU Berlin. <https://doi.org/10.14279/depositonce-4310>.
- Wierstorf, H., M. Geier, A. Raake, and S. Spors. 2013. Perception of focused sources in wave field synthesis. *Journal of the Audio Engineering Society* 61 (1/2): 5–16. <http://www.aes.org/e-lib/browse.cfm?elib=16663>. Accessed 23 Sept. 2019.
- Wierstorf, H., C. Hohnerlein, S. Spors, and A. Raake. 2014. Coloration in wave field synthesis. In *AES 55th International Conference: Spatial Audio, 27–29 August*, Helsinki, Finland, Audio Engineering Society, 1–8. <http://www.aes.org/e-lib/browse.cfm?elib=17381>. Accessed 23 Sept. 2019.
- Wierstorf, H., C. Hold, and A. Raake. 2018. Listener preference for wave field synthesis, stereophony, and different mixes in popular music. *Journal of the Audio Engineering Society* 66 (5): 385–396. <https://doi.org/10.17743/jaes.2018.0019>.
- Wierstorf, H., A. Raake, and S. Spors. 2017a. Assessing localization accuracy in sound field synthesis. *Journal of the Acoustical Society of America*. 141 (2): 1111–1119. <https://doi.org/10.1121/1.4976061>.
- Wierstorf, H., F. Winter, and S. Spors. 2017b. Open science in the Two!Ears project - Experiences and best practices. In *173rd Meeting of the Acoustical Society of America and the 8th Forum Acusticum*. Boston, MA: Acoustical Society of America.
- Wilson, A., and B. Fazenda. 2016. Relationship between hedonic preference and audio quality in tests of music production quality. In *Proceedings of the IEEE 8th International Conference Quality of Multimedia Experience (QoMEX)*, 1–6. <https://doi.org/10.1109/QoMEX.2016.7498937>.
- Winter, F., H. Wierstorf, A. Raake, and S. Spors. 2017. The two!ears database. In *142nd Convention of the Audio Engineering Society*, Berlin, Germany, eBrief 330. <http://www.aes.org/e-lib/browse.cfm?elib=18705>. Accessed 23 Sept. 2019.

- Woodcock, J., J. Francombe, R. Hughes, R. Mason, W.J. Davies, and T.J. Cox. 2018. A quantitative evaluation of media device orchestration for immersive spatial audio reproduction. In *2018 AES International Conference on Spatial Reproduction - Aesthetics and Science*, Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=19606>. Accessed 23 Sept. 2019.
- Zacharov, N. (ed.). 2019. *Sensory Evaluation of Sound*. Boca Raton, FL: CRC Press.
- Zacharov, N., T. Pedersen, C. Pike. 2016a. A common lexicon for spatial sound quality assessment-latest developments. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 1–6. <https://doi.org/10.1109/QoMEX.2016.7498967>.
- Zacharov, N., C. Pike, F. Melchior, and T. Worch. 2016b. Next generation audio system assesement using the multiple stimulus ideal profile method. In *Proceedings of the IEEE QoMEX 2016*, IEEE, 1–6. <https://doi.org/10.1109/QoMEX.2016.7498966>.
- Zahorik, P., D.S. Brungart, and A.W. Bronkhorst. 2005. Auditory distance perception in humans: A summary of past and present research. *ACTA Acustica united with Acustica* 91 (3): 409–420.
- Zieliński, S., F. Rumsey, and S. Bech. 2008. On some biases encountered in modern audio quality listening tests – A review. *Journal of the Audio Engineering Society* 56 (6): 427–451. <http://www.aes.org/e-lib/browse.cfm?elib=14393>. Accessed 23 Sept. 2019.

The Language of Rooms: From Perception to Cognition to Aesthetic Judgment



Stefan Weinzierl, Steffen Lepa and Martin Thiering

Abstract Rooms are not perceptual objects themselves; they can only be perceived through their effect on the presented signal, the sound source, and the human receiver. An overview of different approaches to identify the qualities and the dimensions of “room acoustical impression” will be provided, that have resulted in psychological measuring instruments for room acoustical evaluation from the audience perspective. It will be outlined how the psychoacoustic aspects of room acoustical perception are embedded in a socio-cultural practice that leads to an aesthetic judgment on the quality of performance venues for music and speech.

1 Language and Perception

The aim of this contribution is to highlight the relationship between the characteristics of performance venues for music and speech and the language which is used to describe them. On the one hand, an overview of different approaches to using language as a “measuring instrument” for the qualities of these spaces will be provided. On the other hand, there is an interest in what conclusions can be drawn from the language used with respect to the characteristics of these spaces, the listeners using this language, and the perceptual and cognitive processes involved.

These relationships will be looked at through the lens of a theoretical frame describing the relationship between cultural artifacts (performance spaces), their perception, and their linguistic encoding. This frame model, combining elements of perceptual psychology and cognitive linguistics, assumes a perceptual front end, where an external acoustical signal is transformed into neural activity by auditory sensory organs. For hearing, this process takes place in the inner ear, where sound pressure transmitted by the outer and middle ear is transduced into neural signals. In a perceptual back end, these signals are integrated above different sensory modalities and activate *concepts*, i.e. mental representations corresponding to abstract classes

S. Weinzierl (✉) · S. Lepa · M. Thiering
Fachgebiet Audiokommunikation, Technische Universität Berlin, Berlin, Germany
e-mail: stefan.weinzierl@tu-berlin.de

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_15

of objects, which “tie our past experiences to our present interactions with the world” (Murphy 2004, p. 1) and allow us, for example, to classify an audiovisual experience in a certain social setting as a “concert” and a spatial environment of a specific size, design, and acoustical properties as a “concert hall”. The result of this comparison of sensory information with preconfigured categories (concepts) is called *percept*.

The size and the structure of the concept repertoire as well as the matching process with the sensory input depends on many personal and situational factors, including the knowledge, the experience, the expectations, the motivation, and the attention of the listener. Accordingly, the percept, such as “a successful concert in an acoustically appropriate environment”, depends as much on these situational factors and the conceptual repertoire of the individual listener as it depends on the sensory input at this specific moment.

The relevance of language in this process has two important aspects. First, language is—not the only, but the most important—“metrological” access to human perception. From paired comparisons, similarity judgments, sorting tasks, multidimensional scaling, semantic scales, to vocabulary profiling and related qualitative and quantitative analyses: most studies of the auditory properties of performance venues and sound description in general (Susini et al. 2011) have relied on language-related tasks borrowed from the repertoire of methods of experimental psychology and quantitative and qualitative social research. Second, the language used can itself provide information about the speaker’s conceptual representation of the world (Evans and Green 2007, p. 5), such as about the spatial environments where listeners’ perceive music or speech. The taxonomic organization and the privileged level of categorization that is used in everyday language about spatial concepts are the results of



Fig. 1 Musical events as a cultural, social, visual and acoustic experience. The language to describe performance venues for music and speech reflects each of these domains. The image shows the concept for a new concert hall, to be opened in Munich (©Cukrowicz Nachbaur Architekten)

preceding experiences, and it also shapes the lens through which new experiences are observed. The preferred vocabulary about performance venues does not only reflect the knowledge and the professional experience, e.g., of expert versus lay listeners; it is also assumed to have a direct impact on their instantaneous perception and the respective mental models—for an introduction on mental models see Johnson-Laird (1983), for a description on frame-theory see Minsky (1977). This kind of “linguistic relativism”, which has been evidenced for the languages of different ethnic groups by many empirical observations in cognitive linguistics (Dabrowska and Divjak 2020; Dancygier 2017; Everett 2013; Levinson 2020; Thiering 2018), also occurs on a small scale in lay versus expert language linguistics, or when the languages of groups with different kinds of expertise are compared, such as music listeners versus musicians.

2 Linguistic Inventories as a Basis for Psychological Measuring Instruments in Room Acoustics

Throughout the first half of the 20th century, the investigation of room acoustical environments was mainly focused on the effects of reverberation, with its dependence on frequency, and with its control through volume and absorption according to Sabine’s formula (Sabine 1900). Psychological experiments were conducted already by Sabine himself. In 1902, he invited a number of musical experts to the then recently built New England Conservatory of Music to judge the acoustic quality of piano instruction rooms while seat cushions were successively added to the rooms in order to reduce their reverberation times. Sabine observed that the listeners judged all rooms to be acoustically optimal if the reverberation time was within a quite narrow range of tolerance, from which he concluded a common taste of “*surprising accuracy*” for the acoustical conditions of musical performance venues (Sabine 1906). This unexpected consensus on the appropriate acoustical conditions for classical concert venues, which would hardly have been observed 100 years earlier (Weinzierl 2002), can only be interpreted as the result of a cultural process which accompanied the emergence of public concert life, and which was largely completed around 1900 (Tkaczyk and Weinzierl 2019).

While Sabine asked his subjects only whether the duration of reverberation was appropriate, the British architect and acoustician Hope Bagenal carried out similar experiments with musicians as test participants, who were asked to assess the effect of room acoustic conditions on different sound qualities including “reverberation” (too long/too short), “tone” (full/bright/rich/soft), “tone” (hard/thin/dead/dull), “loudness” (sense of power, body of tone), “reinforcement of notes” (even/uneven) and “conditions”, by which he asked his subjects to name specific halls which resemble the conditions reached (Bagenal 1925). Even if this scheme was not consistently adhered to by his subjects, it can be considered as the first attempt to have the acoustics of concert halls evaluated by a semantic differential, covering the room acoustical effect on spatial, dynamic and timbral dimensions, and considering the effect of typ-

icality with respect to prototypical reference halls. The first standard with guidelines for auditorium acoustics, issued in 1926 by the American Bureau of Standards, however, was mainly focused on specifying an optimal range of reverberation times for halls of different sizes, along with recommendations of how to reach these values (Bureau of Standards 1926).

After 1950, an increasing awareness can be observed that an optimal reverberation time alone is no guarantee for a successful room-acoustical design, and that “reverberance” should not form the only criterion for the perceptual assessment of halls. The British Broadcasting Corporation (BBC) and the Acoustics Group of the Physical Society sponsored a number of experiments, where an identical repertoire (*Don Juan* by R. Strauss) was performed and recorded in four different British concert halls. Subjects listening to the different recordings were then asked to produce a ranking of these halls concerning tonal quality, definition, and overall preference. Consistent rankings, however, could not be observed, and practising musicians exhibited preferences that were different from other skilled listeners (Somerville 1953).

In a further study, a glossary of 14 acoustic terms was collected, which were, according to the authors, commonly used to describe the qualities of concert halls and recording studios—see entry for Somerville and Gilford (1957) in Table 1. A similar list of 18 attributes was proposed by Beranek (1914–2016) in his landmark book on *Music, Acoustics and Architecture*, along with relations between these perceptual qualities and physical properties of the hall, which were based on his intuition and experience (Beranek 1962).

The two lists of attributes, along with a third, originally German list described below, demonstrate the grown awareness for the multidimensional impact of room acoustical conditions on the perceived sound qualities in these halls. At the same time, the studies reflect an awareness for the need to separate the physical and the perceptual domain more clearly. Somerville and Gilford emphasized that the “*subject under investigation is purely aesthetic and therefore must begin and end with human aesthetic judgments*” (Somerville and Gilford 1957, p. 171). Nevertheless, their list is a mix of perceptual and physical items, including aspects such as “scattering” or “standing wave system” without an obvious equivalent in the perceptual domain. Beranek’s features, on the other hand, are psychological throughout, at least if aspects such as “ensemble” and “dynamic range” are understood as “perceived ensemble” and “perceived dynamic range”.

The way to analyze the results of questionnaire studies constructed on the basis of these terms was paved by the psychological fundamentals of the use of semantic differentials (Osgood et al. 1957) and the statistical techniques of multidimensional scaling (MDS, Torgerson 1952) and factor analysis (Spearman and Jones 1950), all of which were introduced during the 1950s. One of the first applications of these new tools was made in a study by Hawkes and Douglas (1971). Sixteen attributes inspired by Beranek’s list, shown in Table 1, were used for a questionnaire applied in four different British Concert Halls (with different musical programs and performers), in the Royal Festival Hall, London, with the newly installed Assisted Resonance system in different technical settings, and in the Royal Festival Hall at 23 different positions. With interest in identifying the different dimensions of acoustic experience (Hawkes

Table 1 Attributes addressed in early investigations on the differential qualities of room-acoustical environments. The translation of the originally German attributes by Wilkens (1977) was adopted from Kahle (1995)

	Somerville and Gilford (1957)	Beranek (1962)	Wilkens (1977)
1	Balance	Intimacy	Small/large
2	Bass masking	Liveness	Pleasant/unpleasant
3	Coloration	Warmth	Unclear/clear
4	Deadness	Loudness of the direct sound	Soft/hard
5	Definition	Loudness of the reverberant sound	Brilliant/dull
6	Diffusion	Definition/Clarity	Rounded/pointed
7	Echoes	Brilliance	Vigorous/muted
8	Flutter echoes	Diffusion	Appealing/unappealing
9	Liveness	Balance	Blunt/sharp
10	Pitch changes	Blend	Diffuse/concentrated
11	Scattering	Ensemble	Overbearing/reticent
12	Singing tone	Immediacy of response	Light/dark
13	Slap back	Texture	Muddy/clear
14	Standing-wave system	Freedom from echo	Dry/reverberant
15		Freedom from noise	Weak/strong
16		Dynamic range	Treble emphasized/not emphasized
17		Tonal quality	Bass emphasized/not emphasized
18		Uniformity throughout the hall	Beautiful/ugly
19			Soft/loud

and Douglas 1971, p. 249), the authors applied both MDS and factor analyses, finding 4–6 orthogonal factors. The solutions they obtained, however, were different both for the different stimulus settings and for the different types of analyses, i.e., both the number of factors and the relation of factors and items were different for each of the sub-studies. This problem will be further addressed in Sect. 3.

While Hawkes and Douglas collected data in the field, i.e., by interviewing concertgoers, Lehmann and Wilkens (1980) used an experimental approach by presenting dummy-head recordings of the Berlin Philharmonic Orchestra in six different halls and capturing the assessment of subjects on a semantic differential with 19 different attributes shown in Table 1. These were selected from a list of originally 27 items by eliminating those with an excessive inter-rater variance, indicating an inconsistent interpretation between subjects. The factor analysis of the ratings delivered three orthogonal factors, explaining 89% of the total variance. Considering the weights of the original attributes on these variables, these were interpreted as *strength*

and extension of the sound source, definition and timbre of the overall sound (Wilkins 1977). In an attempt to overcome the limitations of the experimental approach lacking the ecological validity of live concert situations, Sotirou et al. (1995) used a questionnaire to be rated at three concerts in two different concert halls in London. Similar to Lehmann and Wilkins (1980), they started with a larger vocabulary of about 100 labels, which was reduced to 54 bipolar attributes based on a relevance rating collected in pretests. Analysing the ratings of about 80 participants by factor analysis, the authors obtained four factors explaining roughly 66% of the total variance, which they interpreted as *body*, *clarity*, *tonal quality*, and *proximity*. In both experiments it became obvious that linguistic descriptors are not necessarily suitable as measuring instruments if their meaning is inconsistently interpreted by different raters or if their immediate relationship to the perceptual object under consideration is not assured.

The numerous investigations dedicated to finding suitable *technical* parameters to predict specific perceptual categories are not the subject of this contribution. Only some of them are also interesting here because they highlighted the importance of specific perceptual aspects which did not appear in earlier studies (compare Table 1). Most importantly, a group of studies emphasized the importance of spatial aspects of room acoustics, in particular of an increased perceived “source width” and the perceived acoustic “envelopment” of the auditorium (Barron 1971; Barron and Marshall 1981; Bradley and Soulodre 1995). All of these studies, however, employed synthetic sound fields created by loudspeakers in the anechoic chamber, and asked participants to evaluate these qualities as isolated items. Since they were not evaluated as part of a multidimensional measuring instrument, it is not apparent to what extent they form *independent* aspects of the room acoustical impression, or whether they are physically or perceptually correlated to other aspects.

In this context, it is essential to bear in mind that the ratings of two objects can be correlated because two labels refer to similar perceptual impressions (such as “loudness” and “strength” of sound), or because different perceptual qualities covary in the physical objects of the stimulus pool. For example, rooms providing more “reverberance” could—for physical reasons—always provide more “envelopment”, although the perceptual concepts are clearly different.

After 2000, the study of the perceptual space of room acoustic conditions as a whole attracted a renewed interest directed to the *individual* vocabularies used to describe room acoustic conditions. Several studies, first aiming at the evaluation of spatial-audio reproduction systems (Berg and Rumsey 2006) and then also on the perception of natural acoustical environments, included a qualitative part for the verbal elicitation of the terminology and a quantitative part for the statistical analysis of the generated terms, allowing to identify clusters of attributes with a similar meaning. An initial of two studies conducted at Aalto University, Helsinki, produced room-acoustical stimuli by impulse-response measurements of a loudspeaker orchestra in three different concert halls, encoded in Ambisonics B-Format, processed with directional audio coding (DirAC, Pulkki 2007) and reproduced by a 16-channel loudspeaker system. A second study used impulse responses of eight different concert halls, encoded in Ambisonics B-Format, processed with the spatial impulse-

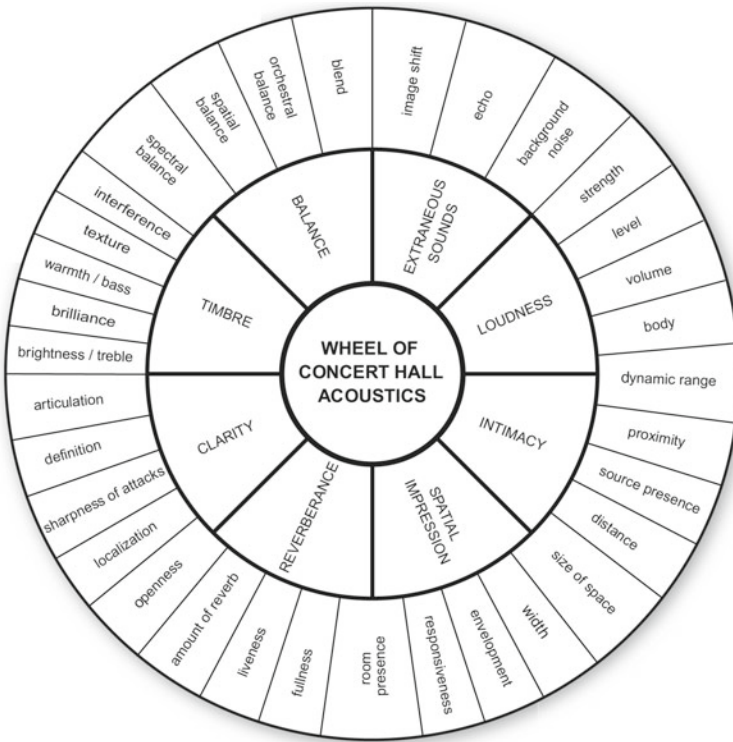


Fig. 2 Wheel of concert hall acoustics (Kuusinen and Lokki 2017)

response-rendering algorithm (SIRR; Merimaa and Pulkki 2005) and reproduced by a 14-channel loudspeaker system. Analysing the large set of about 100 individual attributes generated and rated by 20 resp. 23 participants in the two studies, two resp. three main components could be extracted explaining 66% resp. 67% of the total variance. These were interpreted as *loudness/distance* and *reverberance* in the first study (Lokki et al. 2011), and *loudness, envelopment and reverberance, bassiness and proximity* and *definition and clarity* in the second study (Lokki et al. 2012).

As a summary of their own and other work, authors of the same group suggested a “wheel of concert hall acoustics”, including eight main categories and 33 items to visualize the main perceptual aspects of concert halls with unamplified musical instrument sounds (Kuusinen and Lokki 2017). The wheel format, which has a longer tradition in the domain of food quality and sensory evaluation (compare Noble et al. 1987) is a structured and hierarchical form to present a lexicon of different sensory characteristics. Pedersen and Zacharov (2015) used the wheel to present such a lexicon for reproduced sound, with the selection of the items and the structure of the wheel based on hierarchical cluster analysis and measures for discrimination, reliability, and inter-rater agreement of the individual items—an empirical basis

which was not provided in Kuusinen and Lokki's original wheel for concert halls—see Fig. 2.

3 Psychometrics and Scale Development in Room Acoustics

The studies summarized above have, in different ways, confirmed the multidimensional character of room acoustical conditions as a mediator for the sound qualities of music and speech, and they provided different lists of attributes to describe these qualities. As underlying dimensions of the room acoustical impression, the *loudness* or *strength* and the *reverberance* of rooms were consistently extracted from the ratings of these attributes, as well as a factor for the *timbre* of the room. None of these studies, however, attempted to construct a standardized measuring instrument that can be used with different groups of listeners to describe the whole width of the room-acoustic perception space. From today's point of view, this was impossible due to shortcomings in terms of the experimental stimuli, the participating subjects and the statistical analysis techniques employed in these works. According to modern standards in social research, any psychological measurement instrument has to meet established quality criteria in terms of *reliability*, *validity*, and *invariance* concerning a typical sample of users and stimuli. To establish perceptual dimensions, the investigation has to take care of *representativeness* and *breadth* of stimuli, possible hidden *confounders* and a proper *sample selection* in order to prevent bias concerning the generalizability of results.

A first problem that pertains to most of the early studies in acoustical room impression measurement is the lack of experimental control concerning the stimuli presented. Studies that collected ratings in physically existing rooms always risked the influence of hidden confounders such as the audio content, the visual impression, or the musical performance, all of which co-vary with the auditory impression of the room. Presenting the whole breadth of possible room acoustical conditions while keeping these confounding variables constant seems only possible with state-of-the-art technologies for auralization. It is, of course, true that such an experimental approach can not account for all visual, architectural, and social aspects that constitute the multi-modal impression of a concert venue. To determine the *acoustical* properties of a room, however, these influences act as confounders increasing the measurement error of the test.

One may ask to what extent full control of all non-acoustic factors makes sense, since, for example, an interaction between the room-acoustic conditions and the playing style of musical performers is also present in the real situation and thus not an experimental artifact in the narrower sense. However, incorporating such interaction into the experimental design might, on the one hand, conceal characteristics of space if, for example, musicians were tempted to compensate for the effect of space by adjusting their timbre and volume in the opposite direction. On the other hand, the effect of space on the performer's playing has proven to be quite individual. Different musicians react in very different ways to room-acoustic conditions

(Schärer Kalkandjiev and Weinzierl 2013, 2015; Luizard et al. 2020). The consideration of the interaction of space and performance would, therefore, face the problem of which reaction pattern should be used as a basis here. Nevertheless, the final question that must be taken into account regarding the overall aesthetic judgment of space—“*Is the room suitable for a musical content?*”—cannot be conclusively evaluated in the laboratory.

A second challenge lies in the sample of rooms presented in terms of representativeness. The identification of latent dimensions of room acoustical perception, i.e., a stable factor-analytic solution of the measured data, which is valid beyond the specific sample of rooms used in the test, cannot be expected for a too-small set of stimuli. In order to identify five largely independent perceptual dimensions, a set of at least $2^5 = 32$ stimuli would be required so that all perceptual qualities can be varied systematically and independently from each other and hence can be adequately identified by factor analysis. Furthermore, only with a sufficiently large sample of rooms, the results can be considered representative of the targeted population of room acoustical conditions. Comparing these requirements with the sample sizes used in the studies mentioned above, with typically less than ten rooms, it becomes evident that neither the dimension of the perceptual space, i.e., the number of latent variables, nor the structure and interpretation of the adopted factor solution could be reliably determined.

A third challenge is the size of the sample of listeners. In order to reliably assess the dimensionality of perceptual constructs represented by questionnaire item batteries by factor analyses, it is recommended to use sample sizes of at least 100–200 subjects or at least a sample of three times the number of employed items, even in the favorable case of a good fit of attributes and factors (high commonality). If this requirement is not met—and it never was in the aforementioned studies, stable solutions can be expected only for the most critical factors, i.e., those carrying most of the variance, while all other factors, representing the more subtle aspects of room acoustic impression, are affected by a large sampling error (MacCallum et al. 1999).

Finally, any psychological measuring instrument requires an analysis of the psychometric qualities of the perceptual constructs and questionnaires based on them in terms of validity, reliability, and measurement invariance. Techniques and criteria for this purpose have been developed extensively in the social sciences (Vooris and Clavio 2017). Typical requirements for psychological questionnaire instruments comprise the use of *latent measurement models*, the demonstration of *convergent and discriminant validity* (“Do the scale’s subdimensions actually measure what they are supposed to measure and are sub-dimensions sufficiently different from each other?”), as well as demonstrations of sufficient *construct reliability* (“How precise does the scale measure?”) and *measurement invariance* (Millsap 2011) across time, stimuli, and populations of interest (“Are the scale’s measurements independent of the experimental factors employed?”).

In order to achieve and demonstrate an acceptable degree of validity, reliability, and measurement invariance, and to deal with different sources of measurement error (Schmidt and Hunter 1999), psychological scale development today typically relies on latent-variable models (Loehlin 2004). In this approach, it is assumed that every

manifest measurement of a questionnaire item is, in fact, the expression of underlying latent psychological constructs. When at least three items are measured for any construct exclusively (“simple structure”), it is possible to not only estimate the degree of item-measurement error, construct loading and resulting construct reliability, but also to calculate error-free construct scores and work with these in later analyses. The latent-variable approach also allows checking for reliability across time (retest reliability), which is considered the most important indicator of reliability for scales since Cronbach (1947), and invariance across experimental conditions. Since past studies on room acoustics predominantly drew on principal-component analyses (PCA) and clustering techniques, where none of these tests is possible (Fabrigar et al. 1999), only a minority of reliability-, validity-, and invariance-related questions could be addressed systematically.

4 The Room-Acoustical-Quality Inventory (RAQI)

A multi-stage investigation was conducted by the Audio-Communication Group at TU Berlin to develop a language-based measuring instrument for the different qualities of room-acoustical environments for music and speech and to address the methodological gaps described above (Weinzierl et al. 2018). In a first step, expert knowledge from different professional domains in room acoustics was acquired by help of a focus group in order to provide a comprehensive terminology covering all aspects of the room-acoustical impact on music and speech performances. In a second step, listening experiments with acoustical experts and non-specialists were conducted using 35 rooms of different architectural types, different size, and different average absorption values in order to address the most important types of acoustic performance venues and their specifics. Different audio content was used, including solo music, orchestral music, and dramatic speech. The goals of the subsequent statistical analyses were to,

- Find an exhaustive list of verbal attributes that describes all relevant room acoustical properties
- Identify the best-suited items of this list to form a standardized measurement instrument
- Analyze the underlying dimensions of room-acoustical impressions
- Construct a measurement instrument based on these dimensions and corresponding items
- Demonstrate the reliability of the new instrument across and within raters
- Demonstrate measurement invariance of the new instrument across experimental conditions such as audio content type and subject samples
- Demonstrate sufficient discriminant validity of its subdimensions.

In order to realize this in an experimental setting that permitted controlling for any possible confounders, the study drew on room-acoustical simulation and auralization by dynamic binaural synthesis. The consensus vocabulary generated by the expert focus group consisted of 50 perceptual qualities related to the timbre, geometry,

reverberation, temporal behavior, and dynamic behavior of room-acoustical environments, as well as overall, holistic qualities. While some attributes reflect lower-order qualities closely related to temporal or spectral properties of the audio signal (“loudness”), (“treble/mid/bass range tone color”), perceived “size”, and “width” of sound sources), other attributes reflect higher-order psychological constructs, supra-modal, affective, cognitive, aesthetic, or attitudinal aspects such as “clarity”, “intimacy”, “liveliness”, “speech intelligibility”, “spatial transparency”, or “ease of listening” (Weinzierl et al. 2018, SuppPub 1). For the listening experiment, binaural room-impulse response (BRIR) datasets were simulated for 35 rooms at 2 listening positions for solo music and speech. For the orchestral piece, 25 rooms at 2 listening positions were selected, leaving off 10 rooms where the stage area would not be large enough for an orchestra. Thus, in total, 190 room-acoustical conditions (rooms \times listening positions \times source characteristics) were simulated for the listening experiment. Fourteen of these 190 possible stimulus combinations were rated by each of the 190 participants in a balanced incomplete block design, using 46 items selected from the focus group terminology.

An exploratory factor analysis (EFA) based on the common factor approach was conducted to estimate the number of independent latent dimensions contained in the full-item data matrix. The scree- and Kaiser-criterion was used as a starting point for constructing a multidimensional measurement model. For each of the possible solutions, a series of confirmatory factor analyses (CFA) was conducted to consecutively remove single items from the measurement models up to a point where an implied removal would have led to less than three items per factor, or otherwise, the overall fit of the measurement model was already good. The latter was read from Root-Mean-Square Errors of Approximation (RMSEA), Comparative Fit Indices (CFIs), and Standardized Root-Mean-Square Residual (SRMR) coefficients, as well as congeneric Construct Reliability (CR), indicating the internal consistency of a factor construct, and Average Variance Extracted (AVE) indicating how well a factor explains the scores of its underlying items (Fornell and Larcker 1981).

The factor analysis suggests possible solutions with 4, 6 or 9 factors. These can be interpreted as a general room-acoustical *quality* factor, *strength*, *reverberance*, *brilliance* (4-factor solution), *irregular decay* and *coloration* (6-factor solution), *clarity*, *liveliness* and *intimacy* (9-factor solution). The corresponding item batteries consist of 14, 20, and 29 attributes as shown in Fig. 3. From a statistical point of view, the 6-factor RAQI scale with 20 items is the best compromise between a comprehensive assessment of the full complexity of room-acoustical impressions while at the same time ensuring sufficient statistical independence of the different factors.

With *strength* and *reverberance*, two of the sub-dimensions are omnipresent in the room-acoustical literature. Also, *clarity* and *intimacy* as additional factors have been frequently highlighted by previous studies (Hawkes and Douglas 1971; Lokki et al. 2012). With *brilliance*, *coloration*, and *intimacy* appearing as largely independent factors, it seems that timbre-related qualities play a greater role with more dimensions than previously assumed. The importance of perceived *irregularities in decay* and of *liveliness* as an independent construct has, however, hardly been considered so far.

		Factors	Items	Poles	W	I
9-factor RAQI	6-factor RAQI	Quality	Liking	I like it – I don't like it	1.0	2
			Room acoustic suitability	suitable – not suitable	1.0	56
			Ease of listening	difficult – effortless	0.9	10
			Global balance	balanced – unbalanced	0.8	4
		Strength	Size	small – large	1.0	57
			Loudness	soft – loud	0.7	64
			Width	small – large	0.8	57
		Reverberance	Duration of reverberation	short – long	1.0	47
			Reverberance	dry – reverberant	1.0	54
			Strength of reverberation	weak – strong	1.0	51
			Envelopment by reverberation	weak – strong	0.7	48
		Brilliance	Brilliance	not brilliant – very brilliant	1.0	48
			Tone Color bright/dark	bright – dark	-0.8	-7
			Treble range characteristic	attenuated – emphasized	0.7	3
	4-factor RAQI	Irregular decay	Flutter Echo	none – very strong	1.0	26
			Echo	none – very strong	0.7	40
			Irregularity in sound decay	none – very strong	0.9	32
		Coloration	Boominess	not boomy – very boomy	1.0	37
			Roughness	not rough – very rough	0.7	31
			Comb filter coloration	none – very strong	0.8	34
	Clarity	Temporal clarity	clear – blurred	1.0	10	
		Spatial transparency	blurred – transparent	1.0	1	
		Precision of localization	precise – diffuse	-0.8	-6	
	Liveliness	Liveliness	dead – lively	1.0	11	
		Spatial presence	low – high	1.0	63	
		Dynamic range	small – large	0.9	50	
	Intimacy	Intimacy	remote – intimate	1.0	-3	
		Distance	close – distant	-0.8	51	
		Warmth	cool – warm	0.5	3	
Single items	Metallic tone color	not metallic – very metallic				
	Openness	open – constricted				
	Attack	soft – crisp				
	Richness of sound	low – high				

Fig. 3 The Room-Acoustical-Quality Inventory (RAQI). Four, six, and nine factors as possible sub-dimensions of room acoustical impression can be measured with questionnaires containing 14, 20, and 29 items, which are given with corresponding poles. Weights (W) and Intercepts (I) should be used to measure factors and for structural-equation analysis. Four additional single items with high retest reliability, which could not be assigned to any of the factors, are given below

In terms of psychometric quality, the factors of the 6-factor RAQI exhibit good across-rater consistency and within-rater stability. With regards to measurement invariance, scalar measurement invariance across measurement occasions could be demonstrated for a rather long distance of approximately 42 days. Scores from all RAQI sub-dimensions can thus be directly compared across studies as long as experimental conditions and test subject sample are identical. Similar results pertain to changes in experimental listening position: Scores taken from different listening positions in the same room did not differ systematically. Although the acoustical transfer functions might be quite different, as was demonstrated even for minor changes of the listening position (de Vries et al. 2001), listeners are able to identify the room and its acoustical properties as a consistent cognitive object.

5 Low Retest Reliabilities of Experts versus Laymen: A Problem of Language or a Problem of Perception?

As part of the RAQI development study, the listening test with 88 subjects of 190 in total was repeated six weeks later with identical stimuli. Based on these data, test-retest reliabilities could be determined, calculated as the correlation of measurements within individuals across time, as a measure of the precision, with which certain room-acoustical features could be evaluated. For a majority of the 46 items rated by all participants, the reliabilities turned out to be rather low. Only three items related to reverberation: “*reverberance*”, “*strength*”, and “*duration of reverberation*”, exceeded values of $r = 0.7$, which is usually considered as a criterion of *good* reliability. Many other items, including popular ones in room acoustics such as “*sharpness*” or “*transparency*”, turned out to be based on somewhat unreliable judgements ($r = 0.37/0.43$), using the variation over time within subjects as an indicator.

The low stability of most single item scores indicates that room acoustical impressions appear to be strongly influenced by time-varying situational factors, such as variations in attention, mental efficiency and distraction (*random response errors*) and variations in mood, feeling and mindset (*transient errors*, Schmidt and Hunter 1999). The extent of these psychological measurement errors, however, also depends on the expertise of the listeners. Since the aforementioned subject sample consisted of 60 music-interested non-specialists and 28 individuals with professional education in room acoustics, the relevance of this personal trait could be determined, showing a mean retest reliability of 0.50 across all items for non-specialists versus 0.59 for acoustical experts. Since there is no evidence for differences in the sensory performance between the two groups, the reasons for this difference have to be sought in the perceptual back-end, to pick up on a term from Sect. 1.

Laymen could assess properties, which are clearly room-related and accessible to everyday experience such as the *size* of the room ($r = 0.59$ vs. 0.58 for experts vs. non-experts), the occurrence of *echoes* ($r = 0.67$ vs. 0.68) and even the degree of *liking* ($r = 0.59$ vs. 0.62) with the same, sometimes even better reliability than experts. Whenever, however, the influence of the acoustical source and the influence of the room on the same auditory qualities had to be separated, such as when rating the “*brightness*” ($r = 0.65$ vs. 0.49) and the “*bass-range characteristic*” ($r = 0.64$ vs. 0.17), or the impact of the room on *temporal clarity* ($r = 0.72$ vs. 0.49) and *speech intelligibility* ($r = 0.70$ vs. 0.58), experts were clearly in the advantage. In these cases, the ability to judge this reliably depends on the extent to which *the performance* and *the room* are separated cognitive objects attracting differentiated attention. In addition, experts then also cultivate a specialized vocabulary, including attributes such as *comb-filter coloration* ($r = 0.56$ vs. 0.44) or the *spatial transparency* ($r = 0.56$ vs. 0.38), which are hardly required to describe everyday experiences with music and speech.

Hence, the question of whether the higher precision of experts in evaluating the acoustic properties of rooms is due to a better-trained perception or a more sophisti-

cated vocabulary points to the same interwoven phenomenon: The cognitive performance in the separation of source and space requires a more sophisticated vocabulary, but the sophisticated vocabulary, in turn, can lead to a more differentiated perception.

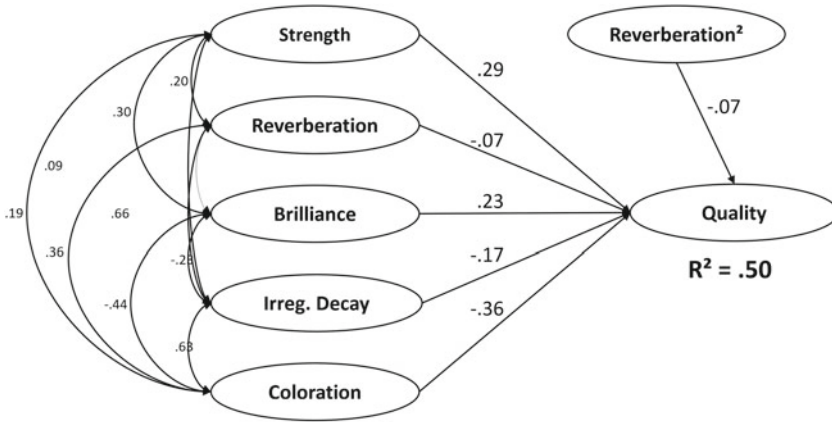
6 The Perceived Quality of Performance Spaces

Most studies dedicated to room acoustic qualities have, in one way or another, also examined what makes up the *overall quality* of performance spaces. To avoid terminological confusion between the two concepts, Blauert has suggested distinguishing between the “aural quality” of a room and a set of “quality features” making up the “aural character” of the room (Blauert 2013). In the following, however, we will stick to a distinction between “quality” and “qualities”, because these are fairly well established terms in the psychological literature.

Quality, in the sense of liking or preference, was already an issue in early investigations aiming at preferred values for the reverberation time of concert halls (Sabine 1906; Watson 1923; Bagenal 1925; Sabine 1928; Knudsen 1931). Multidimensional approaches have often tried to find correlations between the rating of individual attributes or factors and the overall *pleasantness* (Wilkens 1977), the degree of *enjoyment* (Hawkes and Douglas 1971) or the *preference* (Soulodre and Bradley 1995; Lokki et al. 2012) of the room acoustical impression. The relation between the rating of individual qualities and overall quality judgements, however, turned out to be dependent on many factors beyond the room acoustical properties, such as the musical repertoire, the musical performance, and the taste of individual listeners. Some of the studies could even identify different preference groups, one of which preferred more reverberance while the other preferred more clarity and definition (Lehmann and Wilkens 1980; Lokki et al. 2012).

The proportion of the variance in overall quality judgments which can be explained only by acoustic qualities of the rooms themselves was estimated by the authors of this chapter, based on the ratings of 190 participants, collected in the development of the Room-Acoustical-Quality Inventory (Weinzierl et al. 2018). Since an unspecific *quality* factor turned out to be one dimension in each factor solution, the relationship of this general factor to the other dimensions could be estimated. For this purpose, the measurement model of the RAQI was turned into an equivalent structural-equation model regressing the scores of the quality factor on the other factors to estimate their influence and the overall explained variance for the quality factor. Not only linear but also quadratic influences were tested, since psychoacoustic influences on cognitive percepts often show a u-formed (or inverse u-formed) relationship, for example, between preference and reverberation, where an optimal range and a decrease in quality on both sides is the most plausible relation. To account for this, the LMS approach for non-linear effects in structural-equation models was used (Harring et al. 2012).

In the 6-factor version of the RAQI shown in Fig. 3, the model was able to explain about half of the variance in *quality*—see Fig. 4. While *strength* and *brilliance* scores



MLR-Estimation (subject cluster); $\chi^2=715.712$; $df=154$; $p<0.01$; $RMSEA=.037$; $CFI=.956$; $SRMR=.045$

Fig. 4 Structural-Equation Model estimating the influence of five dimensions in the 6-factor-RAQI on the overall quality factor. Path coefficients are beta-weights/correlations. The parameters indicating the fit of the model are explained in Sect. 4

exhibit the largest positive influence on *quality* judgements, a decrease in quality arises with higher *coloration* and higher *irregular decay* values. *Reverberance* had both a linear and an inverted-u relationship to *quality*.

One should be aware that the influence of the individual factors—indicated by the beta weights—as well as the explanatory power of this *quality* model—indicated by the measure of determination, R^2 —depends significantly on the properties of the stimulus pool from which it is derived. The less the presented rooms and the music-and-speech content correspond acoustically to the expectations of the listeners, the higher will be the proportion of the overall quality that can be explained solely by acoustic properties. Also, the sign and the value of the linear beta weights initially only indicate in which direction and to what extent parameters, for example, the rooms’ reverberance, deviated on average from the perceived optimum as seen from the listeners’ point of view. In order to obtain values for these relationships that correspond to a certain cultural practice, it is therefore vital to work with a stimulus pool that is an adequate representation of this practice.

With the stimulus pool used, fifty percent of the variance in *quality* could be explained by perceptual attributes of the room in the presented model. This part of the variance can be considered as the context-independent part of the overall preference judgements, not accounting for the musical repertoire, the musical performance, and the individual taste of the listeners. To explain the preference of music listeners for specific concert halls in a specific situation, a significantly extended model would thus be required. Although various potential influencing factors related to the cultural context of such an overall judgement have been proposed, for example, by Blauert (2013), who pointed out the importance of the *typicality* of concert halls as a result

of two hundred years of Western concert culture, a comprehensive model for the overall aesthetic impression of performance venues, validated by empirical data, has not yet been proposed.

A promising candidate for such a model could emerge when taking into account that judgments about concert halls are always embedded in music-cultural practices in which the music piece, its performance, the performance space and the predisposition of listeners are intricately interwoven, and in their entirety shape the aesthetic judgment of a musical event. Thus, when music psychology examines the factors that influence the aesthetic judgment of a concert performance, the spatial and social context under which that judgment is made will always form a part of that judgment. Listeners can try to consider individual aspects and, for example, try to analytically separate the contributions of the sound source and the performance space to the perceived sound event (Traer and McDermott 2016). However, this separation will always be incomplete. Hence a reasonable approach will possibly lie in applying models that have already been proven empirically to explain the aesthetic judgment of music and music performances also to the evaluation of the *venues* in which music is performed.

Such an approach is exemplified by the model of Juslin et al. (2016), shown schematically in Fig. 5. It assumes that listeners make aesthetic judgments in particular situations in which they adopt what the authors call an “aesthetic attitude”. It is, not least, the concert ritual and the concert hall itself that encourages listeners to adopt this attitude. Once this condition is met, aesthetic processing may be influenced by several factors in the artwork, the perceiver, and the situation. These influences are mediated through the perception, cognition, and emotion of the listener. For one thing, a clear separation of these processes is difficult to draw, and the same musical cues can be processed perceptually (i.e., as sensory impressions), cognitively (i.e., depending on conceptual knowledge) and emotionally (i.e., aroused by other psychophysiological mechanisms). However, even more important is the observation that different listeners use different criteria that determine which of this information and which of those channels have an impact on the resulting aesthetic judgment.

These criteria can be related to varying degrees both to the musical work, to a performance, and to a performance space. One of these criteria is *beauty*. Concerning the performance space, beauty could be understood as the sum of the room acoustical qualities, for which a multidimensional measuring instrument has been developed with the RAQI (Fig. 3). Beauty, however, should not be identified with aesthetic value in general. Other criteria such as the degree of *originality* of a musical event and the related performance venue, the *skill* in its realization, the *typicality* with respect to performance traditions, the degree of *expression* and *emotional contagion*, and the *message* related to the socio-cultural connotations of the musical event, can play a major role for many listeners. In the study of Juslin et al. (2016), most listeners appeared to use a small number of three to five criteria in their judgments, and there were significant individual differences among the listeners, both in how many and which criteria were used.

It is tempting to assume that it is the individual choice of criteria which may account not only for the different aesthetic judgments about music as an integrated

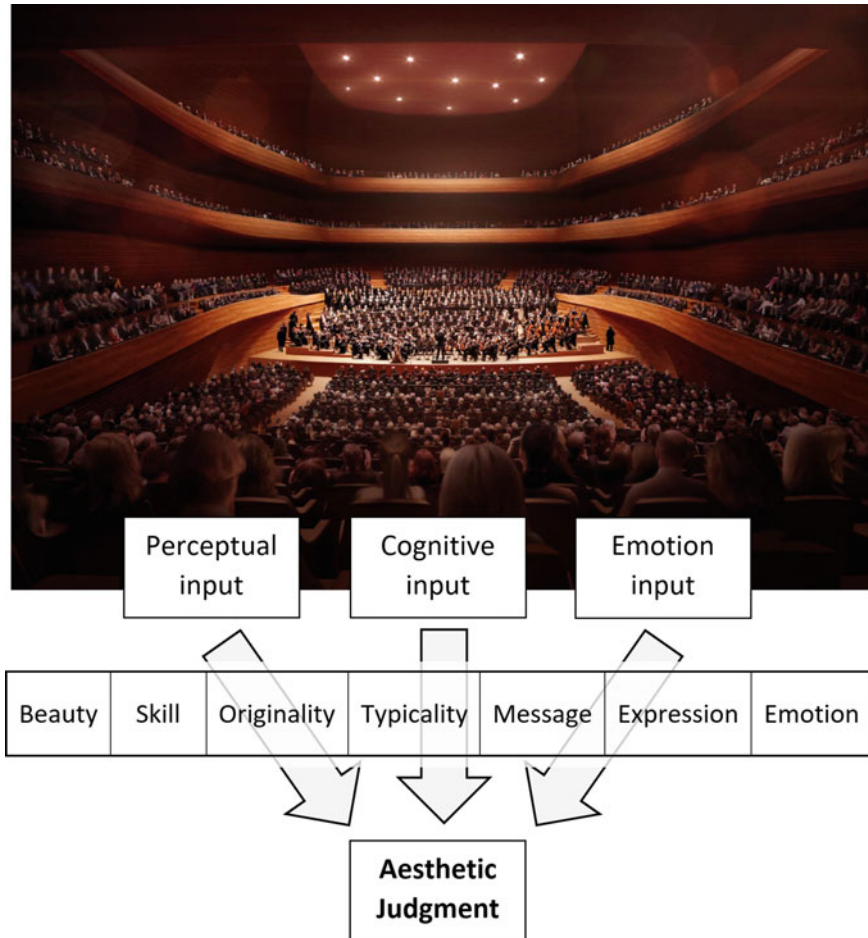


Fig. 5 A model for the formation of aesthetic judgments about music, according to Juslin et al. (2016). The analysis of a musical event is channelled through the perception, cognition, and emotion of the listener. Whether these inputs will affect the resulting aesthetic judgment depends on the listener’s criteria, which act as filters for the processed information

experience but also for the different judgments about musical performance venues. Empirical verification of this assumption could be an essential contribution to a problem that might have been considered for too long only from a psychoacoustic perspective.

Acknowledgements The work reported here was produced within the research unit on “Simulation and Evaluation of Acoustical Environments (SEACEN)”, supported by the Deutsche Forschungsgemeinschaft (FOR 1557). The authors are indebted to all colleagues in this project who contributed to this work, such as D. Ackermann, F. Brinkmann, D. Grigoriev, H. Helmholtz, M. Ilse, O. Kokabi, L. Aspöck, and M. Vorländer, as well as to Jens Blauert for many inspiring discussions on the topic. Further, we want to thank two anonymous reviewers for their comments on this book chapter.

References

- Bagenal, H. 1925. Designing for musical tone. *Journal of the Royal Institute of British Architects* 32 (20): 625–629.
- Barron, M. 1971. The subjective effects of first reflections in concert halls—the need for lateral reflections. *Journal of Sound and Vibration* 15 (4): 475–494.
- Barron, M., and A.H. Marshall. 1981. Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure. *Journal of Sound and Vibration* 77 (2): 211–232.
- Beranek, L.L. 1962. *Music, Acoustics & Architecture*. New York: Wiley.
- Berg, J., and F. Rumsey. 2006. Identification of quality attributes of spatial audio by repertory grid technique. *Journal of the Audio Engineering Society* 54 (5): 365–379.
- Blauert, J. 2013. Conceptual aspects regarding the qualification of spaces for aural performances. *Acta Acustica united with Acustica* 99 (1): 1–13.
- Bradley, J.S., and G.A. Soulodre. 1995. Objective measures of listener envelopment. *Journal of the Audio Engineering Society* 98 (5): 2590–2597.
- Bureau of Standards. 1926. *Circular of the Bureau of Standards, No. 300. Architectural acoustics*. Washington: G.P.O. <https://archive.org/details/circularofbureau300unse>, <https://archive.org/details/circularofbureau300unse>.
- Cronbach, L.J. 1947. Test ‘reliability’: Its meaning and determination. *Psychometrika* 12 (1): 1–16. <https://doi.org/10.1007/BF02289289>.
- Dabrowska, E., and D. Divjak. *Handbook of Cognitive Linguistics*. Berlin, Boston: De Gruyter Mouton.
- Dancygier, B. 2017. *The Cambridge Handbook of Cognitive Linguistics*. Cambridge, UK: Cambridge University Press
- de Vries, D., E.M. Hulsebos, and J. Baan. 2001. Spatial fluctuations in measures for spaciousness. *Journal of the Acoustical Society of America* 110 (2): 947–954.
- Evans, V., and M. Green. 2007. *Cognitive Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.
- Everett, C. 2013. *Linguistic Relativity: Evidence Across Languages and Cognitive Domains*, vol. 25. Berlin/New York: De Gruyter Mouton.
- Fabrigar, L.R., D.T. Wegener, R.C. MacCallum, and E.J. Strahan. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods* 4 (3): 272–299.
- Fornell, C., and D.F. Larcker. 1981. Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research* 18 (1): 39–50. <https://doi.org/10.1177/002224378101800104>.
- Harring, J.R., B.A. Weiss, and J.C. Hsu. 2012. A comparison of methods for estimating quadratic effects in nonlinear structural equation models. *Psychological Methods* 17 (2): 193–214. <https://doi.org/10.1037/a0027539>.
- Hawkes, R.J., and H. Douglas. 1971. Subjective acoustic experience in concert auditoria. *Acustica* 24 (5): 235–250.
- Johnson-Laird, P.N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge: Harvard University Press.
- Juslin, P.N., L.S. Sakka, G.T. Barradas, and S. Liljeström. 2016. No accounting for taste?: Idiographic models of aesthetic judgment in music. *Psychology of Aesthetics, Creativity, and the Arts* 10 (2): 157–170.
- Kahle, E. 1995. *Validation d'un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras (Validation of a perceptual model of the acoustic quality in an ensemble of concert halls and opera houses)*. Le Mans: Le Mans Université. Ph.D. thesis.
- Knudsen, V.O. 1931. Acoustics of music rooms. *Journal of the Acoustical Society of America* 2: 434–467.
- Kuusinen, A., and T. Lokki. 2017. Wheel of concert hall acoustics. *Acta Acustica united with Acustica* 103 (2): 185–188.

- Lehmann, P., and H. Wilkens. 1980. Zusammenhang subjektiver Beurteilungen von Konzertsälen mit raumakustischen Kriterien (Relation between subjective evaluations of concert halls and room-acoustical criteria). *Acustica* 45: 256–268.
- Levinson, S.C. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.
- Loehlin, J.C. 2004. *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*. Mahwah: Routledge.
- Lokki, T., J. Pätynen, A. Kuusinen, and S. Tervo. 2012. Disentangling preference ratings of concert hall acoustics using subjective sensory profiles. *Journal of the Acoustical Society of America* 132 (5): 3148–3161.
- Lokki, T., J. Pätynen, A. Kuusinen, H. Vertanen, and S. Tervo. 2011. Concert hall acoustics assessment with individually elicited attributes. *Journal of the Acoustical Society of America* 130 (2): 835–849.
- Luizard, P., J. Steffens, and S. Weinzierl. 2020. Singing in different rooms: Common or individual adaptation patterns to the acoustic conditions? *Journal of the Acoustical Society of America*. 147 (2): EL132–EL137.
- MacCallum, R.C., K.F. Widaman, S. Zhang, and S. Hong. 1999. Sample size in factor analysis. *Psychological Methods* 4 (1): 84–99.
- Merimaa, J., and V. Pulkki. 2005. Spatial impulse response rendering I: Analysis and synthesis. *Journal of the Audio Engineering Society* 53: 1115–1127.
- Millsap, R.E. 2011. *Statistical Approaches to Measurement Invariance*. New York: Routledge.
- Minsky, M. 1977. Frame-system theory. In *Thinking. Readings in Cognitive Science*, ed. P.N. Johnson-Laird and P.C. Wason. Cambridge: Cambridge University Press.
- Murphy, G. 2004. *The Big Book of Concepts*. MIT Press.
- Noble, A.C., R.A. Arnold, J. Buechsenstein, E.J. Leach, J.O. Schmidt, and P.M. Stern. 1987. Modification of a standardized system of wine aroma terminology. *American Journal of Enology and Viticulture* 38 (2): 143–146.
- Osgood, C.E., G.J. Suci, and P.H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana, Ill.: University of Illinois Press.
- Pedersen, T.H., and N. Zacharov. 2015. The development of a sound wheel for reproduced sound. *Audio Engineering Society Convention*. 138 Preprint No. 9310.
- Pulkki, V. 2007. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society* 55: 503–516.
- Sabine, P.E. 1928. The acoustics of sound recording rooms. *Transactions of the Society of Motion Picture Engineers* 12 (35): 809–822.
- Sabine, W.C. 1900. Reverberation. *The American Architect and Building News*. 68: 3–5, 19–22, 35–37, 43–45, 59–61, 75–76, 83–84.
- Sabine, W.C. 1906. The accuracy of musical taste in regard to architectural acoustics. *Proceedings of the American Academy of Arts and Sciences* 42 (2): 53–58.
- Schärer Kalkandjiev, Z., and S. Weinzierl. 2013. The influence of room acoustics on solo music performance: An empirical case study. *Acta Acustica united with Acustica* 99: 433–441.
- Schärer Kalkandjiev, Z., and S. Weinzierl. 2015. The influence of room acoustics on solo music performance: An experimental study, Psychomusicology: Music. *Mind and Brain* 25 (3): 195–207.
- Schmidt, F.L., and J.E. Hunter. 1999. Theory testing and measurement error. *Intelligence* 27 (3): 183–198. [https://doi.org/10.1016/S0160-2896\(99\)00024-0](https://doi.org/10.1016/S0160-2896(99)00024-0).
- Somerville, T. 1953. Subjective comparisons of concert halls. *BBC Quarterly* 8: 125–128.
- Somerville, T., and C.L.S. Gilford. 1957. Acoustics of large orchestral studios and concert halls. *Proceedings of the IEE* 104: 85–97.
- Sotiropoulou, A.G., R.J. Hawkes, and D.B. Fleming. 1995. Concert hall acoustic evaluations by ordinary concert-goers: I, Multi-dimensional description of evaluations. *Acta Acustica united with Acustica* 81 (1): 1–9.

- Soulodre, G.A., and J.S. Bradley. 1995. Subjective evaluation of new room acoustic measures. *Journal of the Audio Engineering Society* 98 (1): 294–301.
- Spearman, C., and L.W. Jones. 1950. *Human Ability*. London: Macmillan.
- Susini, P., G. Lemaitre, and S. McAdams. 2011. *Psychological Measurement for Sound Description and Evaluation in Measurement with Persons: Theory, Methods, and Implementation Areas*, eds. B. Berglund and G. B. Rossi. New York: Psychology Press.
- Thiering, M. 2018. *Kognitive Semantik und Kognitive Anthropologie: Eine Einführung [Cognitive semantics and cognitive anthropology: An introduction]*. Berlin: De Gruyter.
- Tkaczyk, V., and S. Weinzierl. 2019. Architectural acoustics and the trained ear in the arts: A journey from 1780 to 1830. In *The Oxford Handbook of Music Listening in the 19th and 20th Centuries*, eds. C. Thorau and H. Ziemer. New York, NY: Oxford University Press.
- Torgerson, W.S. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika* 17 (4): 401–419.
- Traer, J., and J.H. McDermott. 2016. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences* 113 (48): E7856–E7865.
- Vooris, R., and G. Clavio. 2017. Scale development. In *The International Encyclopedia of Communication Research Methods*, eds. C.S.D.J. Matthes and R.F. Potter. American Cancer Society.
- Watson, F.R. 1923. *Acoustics of Buildings*. New York: Jon Wiley and Sons.
- Weinzierl, S. 2002. *Beethovens Konzerträume. Raumakustik und symphonische Aufführungspraxis an der Schwelle zum modernen Konzertwesen [Beethoven's concert halls. Room acoustics and symphonic performance practice on the threshold to modern concert life]* (Bochinsky, Frankfurt am Main).
- Weinzierl, S., S. Lepa, and D. Ackermann. 2018. A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI). *Journal of the Acoustical Society of America* 144 (3): 1245–1257.
- Wilkens, H. 1977. Mehrdimensionale Beschreibung subjektiver Beurteilungen der Akustik von Konzertsälen [Multidimensional description of subjective evaluations of the acoustics of concert halls]. *Acustica* 38: 10–23.

Modeling the Aesthetics of Audio-Scene Reproduction



John Mourjopoulos

Abstract Reviewing work from diverse scientific fields, this chapter approaches the human aesthetic response to reproduced audio as a process of attraction and efficient (“fluent”) processing for certain auditory stimuli that can be associated with listener pleasure (valence) and attention (arousal), provided that they conform to specific semantic and contextual principles, either derived from perceived signal features or from top-down cognitive processes. Recent techniques for room-related loudspeaker-based presentation of auditory scenes, especially via multichannel reproduction, further extend the options for manipulating the source signals to allow the rendering of virtual sources beyond the frontal azimuth angles and to enhance the listener envelopment. Hence, such methods increase arousal and valence and contribute additional factors to the listeners’ aesthetic experience for reproduced natural or virtual scenes. This chapter also examines the adaptation of existing models of aesthetic response to include listeners’ aesthetic assessments of spatial-audio reproduction in conjunction with present and evolving methods for evaluating the quality of such audio presentations. Given that current sound-quality assessment methods are usually strongly rooted in objective, instrumental measures and models, which intentionally exclude the observers’ emotions, preferences and liking (hedonic response), the chapter also proposes a computational model structure that can incorporate aesthetic functionality beyond or in conjunction with quality assessment.

1 Introduction

1.1 Overview

This book chapter reviews material related to the aesthetics of recorded sounds, with the emphasis on the spatial relationships of sound objects and scenes represented via

J. Mourjopoulos (✉)

Wire Communications Laboratory, Audio and Acoustic Technology Group, Electrical and Computer Engineering Department, University of Patras, 26500 Patras, Greece
e-mail: mourjop@upatras.gr

© Springer Nature Switzerland AG 2020

J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_16

455

current loudspeaker and room-related reproduction technologies. The review draws evidence from many scientific disciplines, such as audio technology, acoustics, neurosciences, cognitive psychology, and philosophy, to highlight underlying convergent trends in scientific research that are relevant to the aesthetic awareness of recorded-reproduced sound. Note that in this chapter, as “recorded-reproduced sound” all physical and man-made processes are considered that combine pre-recorded or synthetic audio data into loudspeaker-based (“room-related”) presentations of auditory scenes via such loudspeaker arrangements (Toole 2018; Blauert et al. 2013; Rumsey 2017; Breebaart and Faller 2007).

Human appreciation of sound qualities, auditory scenes, and music, must be considered as an unprecedented phenomenon of biological functionality of the auditory periphery, the brain, and cognition, to the end of analyzing, organizing, classifying and combining acoustic sensory stimuli. Such unique coding and recombination of information are affecting numerous centers in the brain and, clearly, proper analysis of these phenomena is beyond the scope of this chapter—compare Levitin (2011), Lund and Mäkivirta (2017), McDermott (2012). Instead, here an engineering perspective is followed, that is, such complex perceptual, cognitive, and biological processes are approached via functional models that consider human aesthetic response as a mediating mechanism between perception and interpretation. Such mediating mechanism is considered to function in series, in parallel, or even independently from any related quality judgment initiated by the same stimuli.

The complexity of modeling human aesthetic response is immense and interdisciplinary, quoting Julien P. Renoult: “...*Few topics can take pride in transcending the traditional frontiers between disciplines from the humanities and the sciences as much as aesthetics...*” (Renoult 2016). Furthermore, it remains virtually unexplored in the field of audio and acoustics. In contrast, the aesthetic analysis of various art forms including music has been a well-established area of the fields of humanities, philosophy and musicology (Wikipedia 2018; Brattico et al. 2017). Any such analysis of musical aesthetics and aspects related to the music content will be carefully excluded from this book chapter following instead an engineering perspective and drawing evidence from similar current research in representational aesthetics of visual arts and images (Joshi et al. 2011; Deng et al. 2017).

The aesthetics of listening to recorded-reproduced music can be seen as a hierarchical representation of mental abstractions and emotional response to audio stimuli exceeding a specific threshold of attention (Brattico et al. 2013, 2017). Cognitive representations at this level of aesthetic and emotional response are not easily represented via verbal descriptors or objective classification and, hence, are not easily analyzed via the established psychometric methods as usually applied to audio-sound-quality assessment. Hence, the aesthetic response must be considered in conjunction and beyond existing instrumental, descriptive or computational methods for judging sound/audio qualities, even if, as will be noted later, there is often some overlap between quality and aesthetic judgments. Given the significant body of engineering methods in the field of audio-quality assessment (Bech and Zacharov 2006; Raake and Wierstorf 2014; Zacharov et al. 2016; Pedersen and Zacharov 2015), this

book chapter will also consider respective relationships between aesthetic models and methods for evaluating sound quality.

Recent techniques for room-related (loudspeaker-based) or head-related (headphone-based) presentation of auditory scenes, especially when optimized via digital signal processing (Rumsey 2017; Breebaart and Faller 2007; Pulkki et al. 2018), further extend the options for manipulating the perceived spatial scenes often aiming beyond authenticity and realism in their presentations. Such techniques often intentionally enhance listener inhibition and envelopment hence affecting the listener interpretation of the recorded or synthesized auditory scenes. It is clear that today, as for all periods since the inception of recording technology, there are diverse approaches to the technical and aesthetic rules for modifying or synthesizing recorded sounds, that is, the degree of the imposed illusion and impact designed into such representations compared to natural listening.

Historically, such semi-autonomous reference system between reality and its virtual representation as widely used for more than 100 years (Toole 2018; Hamilton 2003), must be considered as the “grandfather” of the more recent virtual-reality (VR) technologies. Given the multiplicity of largely unexplored facets of the problem, this chapter specifically considers the aesthetic implications of the techniques that allow the representation of spatially separated sound-music objects and their relationships to a realistic or virtual sound field. In particular, this chapter presents

- A literature review of the functional models of aesthetic response drawing evidence from current research in the fields of computer vision, image analysis as well as from cognitive psychology.
- An analysis of audio-technology developments relevant to the perceived qualities and aesthetics of the reproduced auditory scenes.
- An analysis of signal and system features that affect aesthetic judgment beyond or in conjunction to sound-quality evaluation of such recorded and reproduced sounds also discussed in detail in Raake and Wierstorf (2020), this volume.
- A proposal for the structure of a functional engineering model combining perceptual and cognitive mechanisms related to aesthetic judgments for recorded sounds and music, focusing on the spatial parameters of the audio reproduction technology.

1.2 Problem Formulation

This chapter focuses on the aesthetics of the technical medium of sound-reproduction representation. As much as possible it is thereby avoided to consider the aesthetics of music content, context, composition, form, performance, style, etc. A selection of statements regarding current listening trends is listed in the following.

- Today almost all music reaching the listeners is produced, recorded and reproduced electro-acoustically and, hence, such according listening experience has become ubiquitous to everyone (Hamilton 2003; Rumsey 2008)

- With the widespread exposition to recorded sounds for more than a century, audio-engineering practices form an integral aspect of the experience of music appreciation for most humans over a variety of situations, environments. The recorded sound is delivered over highly diverse equipment and conditions
- Virtually all original music performances are manipulated in the electronic-digital domain to achieve an acceptable level of technical and aesthetic perfection. As Hamilton has pointed out (Hamilton 2003), recording has transformed the nature of music as an art by reconfiguring the cognitive hierarchy of the aesthetics of perfection and imperfection
- Due to the varying degree of signal manipulation, it is impossible for most musical genres to use as reference an original source or auditory scene. Thus judgments on reproduction authenticity and transparency can never be fully objective—especially considering additional and case-dependent artifacts imposed during reproduction (Brattico and Pearce 2013; Rumsey 2008)
- Listening impact, plausibility, realism, and arousal from recorded sounds can challenge the level of the emotional effect achieved via natural listening (Eerola 2014). This is evident from the wide acceptance and viability of recorded-music media in social, artistic, cultural and economic terms. Recorded music has thus gained profound significance for each one of us (Brandenburg et al. 2020; Rumsey 2002)
- Recorded music has eventually established a novel global framework for aesthetic representations of sound and musical objects (Brattico et al. 2017; Hamilton 2003). As Kahn writes “... *Phonography challenges music’s hegemony as universal art of sound...*” (Kahn 2001)
- Thus, it is now accepted that any technical shortcomings in real-life events and in performances is to be corrected to achieve listener acceptance via perceptual plausibility and, therefore, also to be engineered via autonomous aesthetic rules and concepts
- Such aesthetic aspects are highly significant to specialists such as musicians, music educators, audio engineers, experienced listeners/audiophiles. Functional modeling of these aesthetic aspects can be applied to enhance the scope of existing sound-quality assessment methods
- Due to their lengthy exposure, listeners in modern societies tend to accept audio production and audio reproduction as a common cultural artifact that obeys its own semantic rules (Rumsey 2011)
- These rules are often stylized and “group-specific”. Consequently, cultural bonding within listeners is highly active. This increases the dependence of the listeners’ judgments as regards the analysis of individual preferences or of cultural/social biasing (Brattico and Pearce 2013; Zielinski et al. 2008)
- Current and emerging audio technologies provide extended control of the sound field and the properties of natural or artificial sources, along with their spatial representation to the listener. Hence, spatial aspects of the audio technology contribute increasingly to current and future aesthetic experience (Rumsey 2002, 2017)
- “Head-related” (via headphones) and “room-related” (via loudspeakers) reproduction of auditory scenes often follow different technical solutions and suffer from different constraints (Rumsey 2017).

1.3 Encoding and Decoding

Focusing now on room-related auditory-scene recording/reproduction via loudspeakers, the well-established studio techniques that form the foundations of acoustically-transparent encoding procedures will not be considered at this point. Instead, for simplicity, the focus is laid on cases where the original material has been prerecorded into discrete audio channels, and where, during the encoding, the components are synthetically combined into an intended auditory scene. Given the ever-increasing options for format encoding and reproduction (e.g., stereophonic, surround, holophonic), (Rumsey 2015b, 2017), this simplification may exaggerate the aesthetic options offered to contemporary sound engineers and musicians via appropriate software tools. As noted by Rumsey (2002), “...*The computer music composer, along with most studio audio engineers will be faced with novel and unknown practices when attempting to explore the possibilities offered by the (new) multichannel audio formats...*”. Rumsey categorizes the aesthetic options offered for such encoding, as allowing to

1. *Create virtual space*: Redefine the composition-parameter hierarchy to include space, retain musical coherence and focus, control relationship between musical substance and effect
2. *Assign spatial roles*: Define spatial roles for sources, decide for themes/sources assigned to the central position, assign musical significance to image width and/or movement, explore spatial aspects of timbre
3. *Explore extra-musical aspects*: Relate spatial audio imagery to visual action (if visual data are also used), define listener perspective, explore plots in the extended audio soundscape.

With respect to the options of audio reproduction, the current high-quality room-related loudspeaker methods for the presentation of auditory scenes offer a wide battery of software tool, in particular, when these scenes are synthesized (Rumsey 2009). Briefly, two state-of-the-art methods aim at physically-authentic 3D-sound representation (“Holophony”), namely, Higher-Order Ambisonics (HOA) and Wave-Field Synthesis (WFS). Each of them achieves a different level of acoustic 3D-performance, especially with respect to scalability and the capability for adaptations to the loudspeaker-room reproduction system, whereby each suffers from well-known limitations (Rumsey 2009). The specific deficits of each of the methods can be compensated up to certain a degree by additional encoding-preprocessing (e.g., via amplitude panned or VBAP encoded channels, binaural parametric-cue selection, system-response compensation)—compare Rumsey (2009), Pulkki (2001), Bertet et al. (2006), Merimaa and Pulkki (2005), Faller (2004), Daniel et al. (2003).

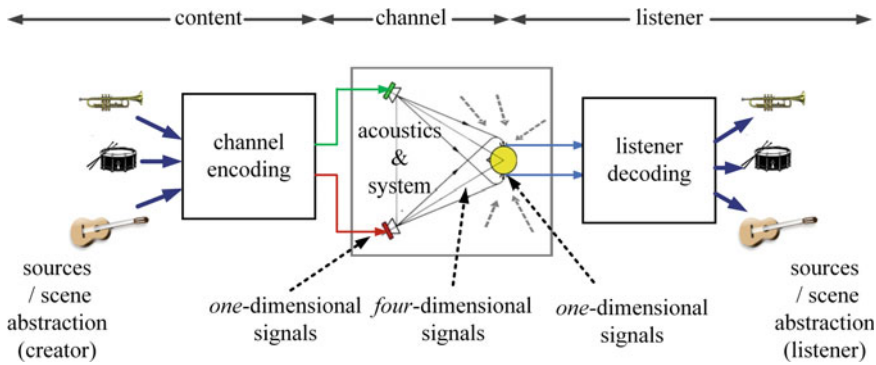


Fig. 1 Conceptual general representation of a complete transmission channel, incorporating recording and room-related reproduction

1.4 The Transmission Channel

Irrespective of the encoding/decoding methods, the listeners will respond and provide evaluations for sound-quality and aesthetics related to the reproduced sound event and scene as is modified via the specific transmission channel (Toole 2018; Bech and Zacharov 2006; Rumsey 2002, 2017). In general terms the complete end-to-end chain from “input” to “output” may be approached conceptually as a transmission system, incorporating the sound engineer’s choices for encoding and preprocessing, the electroacoustic-loudspeaker reproduction, room acoustic response effects and, finally, the listeners auditory, perceptual, and cognitive decoding.

A conceptually simplified representation of such a system, which comprises the end-to-end chain from the source to the receiver, is shown in Fig. 1. For an extended analysis of the end-to-end sound-processing chain compare (Raake and Wierstorf 2020), this volume. In such a system, it is evident that aesthetic judgements can be affected, applied and related to *each individual step and component* of the chain as well as to the global system response. On the left side of Fig. 1, the conceptual abstraction in the brain of the sound engineer or musician for synthesizing the original acoustic scene is, typically realized by downmixing any number of source channels (tracks) into, say here for simplicity, 2-channel stereo mix, using appropriate choice of processors and simple panning-law parameters. For this discussion, such input signals are not “real”, that is, recorded via microphones following established principles for capturing sources within acoustic spaces, but instead are they are pre-recorded, processed or electronically-generated signal tracks.

For the generation of the desired spatial relationships between the signals of the individual tracks, the sound engineer relies mostly on panning, that is, adjusting the relative amplitude in such a way that each track is assigned either to the left or the right channel of the downmixed stereo signal. Such relative amplitudes of the specific track-signals will be retained over to the reproduction stage. Thus, it generates the desired perceived image of this source in the appropriate direction

between the loudspeakers. Additional spatial cues may be introduced to the individual tracks, for example, regarding artificial reverberation, equalization and delays, aiming at synthesizing the impression of variable direct-reverberant ratios in the signal within the stereo perspective. Such features are impressed on the stereo downmix signal. Hence, they are presented as input to the transmission channels during room-related reproduction—Fig. 1. Listening to any of the individual tracks of a musical piece in isolation will typically not lead to the intended positive aesthetic evaluation and may not correspond to the desired effect of scene abstraction. However, the combined processed downmixed sum of these elements is capable of rendering an agreeable aesthetic level (Rumsey 1998; Brattico et al. 2017).

The downmixed source material is usually further manipulated by a mastering engineer. This specialist imposes technical and aesthetic rules on the encoded (e.g., stereo) data (Rumsey 1998; Katz 2015). Such processing is typically carried-out under acoustically-controlled reproduction conditions via highly accurate audio equipment—often with informal comparative listening over different systems. Irregularities or other factors reducing its aesthetic appeal are suppressed this way. Several quality-relevant features have been found to statistically correlate with the manipulated mastered signal, predominantly an increased loudness (Katz 2015; Vickers 2010).

The room-related reproduction of this material is taking place in a different space than in the mixing room, typically in the listener’s own room. For the case of listening to stereo reproduction, each of two loudspeakers reproduces a 1-dimensional acoustic signal into the listening space. Such signals spread via a complex 3D spatial pattern, largely depending on the loudspeakers on-axis responses, their directivity, and their placement. Broadly speaking, they generate the direct-path transmission plus multiple reflections. In other words, they generate four-dimensional signals, received as binaural 1-dimensional signals at each of the listener’s ear-canal entrances—Fig. 1. In this section, it is avoided as much as possible to consider the aesthetics of musical content, context, composition, performance, style, form, performance, style, and related items. The combined room-related effects of the transmission subsystems are shown in Fig. 2. Any aesthetic assessment of non-auditory effects may be also

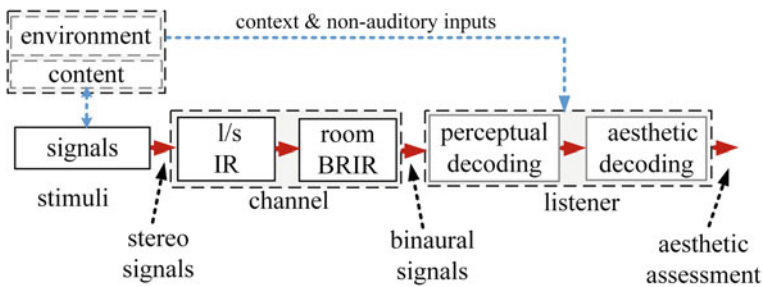


Fig. 2 Components of the room-related reproduction channel along with a conceptual model of aesthetic assessment, eventually performed by a virtual assessor

conceptually depicted in the model of Fig. 2¹—if available. For any specific source (loudspeaker), room and listener position, the combined effects of the transmission channel on the original signals up to the entrance of each left and right ear canal can be represented via the corresponding *binaural room-impulse responses* (BRIR). From these responses, perceptually-relevant parameters such as spectral alterations (coloration), acoustic reverberation and interaural differences can be extracted. A formal analysis of the relationship between the physically measurable BRIRs and the resulting auditory-relevant functions is provided in Grosse and Par (2015).

For the cases discussed, the received signals are greatly modified, that is, distorted in all physical domains by the complex interactions between the loudspeakers' on-axis/off-axis linear responses, by electroacoustic chain non-linearities, room acoustics, the relative placement of listeners and loudspeakers within the space, and so on (Toole 2018; Volk et al. 2017). Such distortions clearly affect the reproduced spectral overall balance, namely, the signal timbre (i.e. coloration and phase delays) (Toole 2018; Smith and Bocko 2017), the time-domain evolution (especially signal attacks), and the dynamic range. Specifically-encoded source features are thus potentially less distinct and hence the qualitative and aesthetic targets as intended by the creators/engineers are affected in an unpredictable—in most cases negative way. As is discussed in detail in Raake and Wierstorf (2020), this volume, “...*beyond reproducing the physically correct sound pressure at the ear drums, more effects play a significant role in the quality of the auditory illusion. In some cases, those can dominate the perception and even overcome physical deviations...*”. Significant degradations are also generated as regards the spatial accuracy of the reproduced signals. These degradations cannot be accessed via typical acoustical in situ measurements since they depend on many variables related to loudspeaker and listening-space interactions (Volk et al. 2015, 2017). Such aspects are even more complex when holophonic scene presentation is employed—compare Nicol (2020), this volume.

The most prominent qualitative aspects for such a spatial interpretation of the intended scene are source-image localization, sweet-spot robustness (for varying listener placement relative to loudspeakers), auditory spaciousness (listener envelopment and source width)—Toole (2018), Volk et al. (2017), Francombe et al. (2015, 2017, 2018), Mason (2017), Lepa et al. (2014). Given that acoustic evaluation of spatial qualities is very difficult and it is strongly dependent on binaural perception, methods have been proposed that employ a binaural-parameter extraction, analysis and classification approach for typical stereo and surround sound (e.g., 5.1) setups in different listening-room scenarios (Kamaris and Mourjopoulos 2018; Wierstorf et al. 2013a) but also for assessing the accuracy of holophonic presentations Wierstorf et al. (2013a), Grosse and Par (2015) and Nicol (2020), this volume.

An example of a spatial map for stereo image localization accuracy evaluated by such a “virtual listener” binaural model is shown in Fig. 3. The model is driven by source signals of short noise bursts, panned along all potential angles between the stereo loudspeakers, and transmitted to any potential listener positions along a

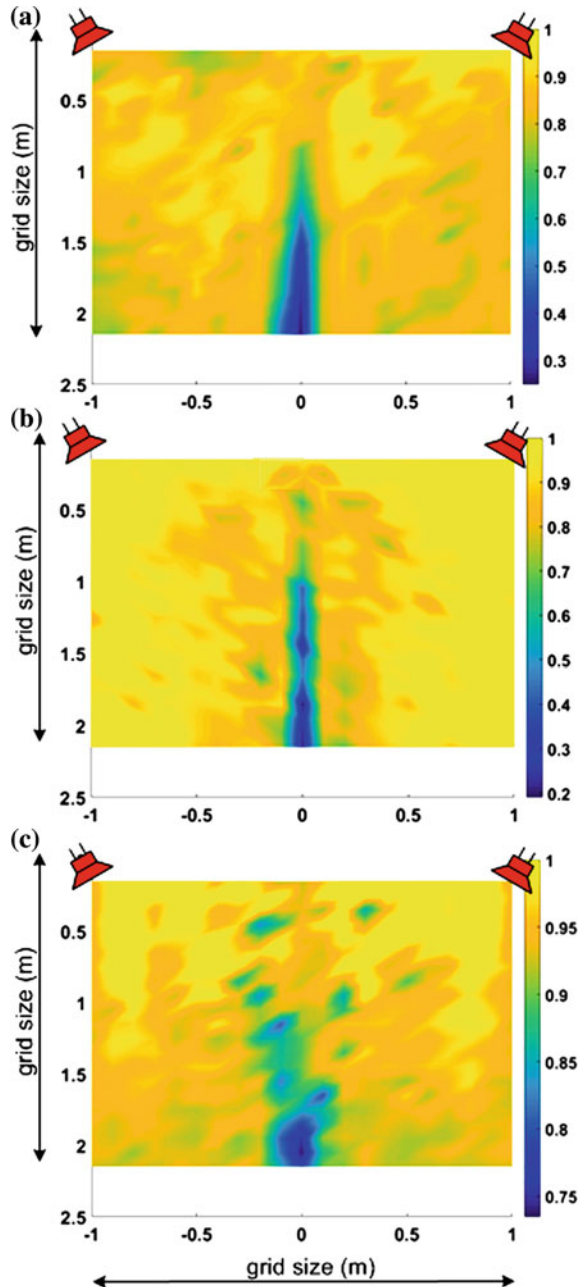
¹The listeners also receive cross-modal information, for instance, a visual one, and may draw from prior knowledge of the scene, that is, they use cognition.

plane grid. Such transmission is enabled by use of the corresponding BRIRs for the reproduced signals before they reach the ear canal entrance at each listener grid position. Thus, the BRIRs combine all effects of the specific loudspeaker frequency response and directivity and the room reflections with respect to the listener positions. The binaural virtual-listener model analyses the perceptually-relevant spatial cues (e.g., ITDs and ILDs) for the evaluation of the perceived *directions-of-arrival* (DOAs) of the reproduced signals. These directions are then compared to the intended DOAs for the specific image panning with respect to the listener in the grid. The potentially perceived errors in the DOAs due to channel distortions are color-mapped along the grid, indicating the area of best localization.

The *sweet spot* should correspond to a well-defined symmetric strip along the middle axis of the loudspeakers. As can be seen in Fig. 3, even small room asymmetry and reverberation can blur the area of best localization accuracy. After perceptual decoding in the listeners' auditory system and brain, the two input signals to the ear canals are transformed into a perceived abstraction of the original acoustic scene, that is, into a perceived auditory event and auditory objects in the listeners' perceptual world. From an engineering perspective, a common goal is that acoustic sound sources and the associated auditory events and spatial-scene abstractions, must be "identical", that is, authentic reproductions. It is hypothesized that this can be achieved with complete transmission channels that are distortion-free, that is, "transparent" (Rumsey 2002). And indeed, this physicalistic hypothesis carries a long way when engineering spatial-audio system.

However, the concept of transparent transmission has its limitations when perceptual and cognitive effects are included in the discussion. Thus, the notion of perceptually-viable ideal transmission needs to be extended from the mathematical definition of *ideal*. This introduces a difficulty in specifying the *ground truth* or transparency condition for any subsequent quality analysis of any specific listening scenario. This problem will be discussed in more detail in Sect. 2.4. A broader definition of "ideal" in this context refers mostly to plausibility judgements. Note that for various reasons there is no direct reference to the original ground-truth reference. Furthermore, due to the complex loudspeaker-room spatial coupling with such scenes, non-auditory-modality inputs, semantic and contextual factors, and even due to differences in the source material and signal properties, the quality and aesthetic appraisal lacks "any ground-truth" assessment—compare Rumsey (2002). At large, sound-quality and aesthetic judgements must be regarded as the degree to which any auditory event fulfills expectations. Such comparisons need to be performed against a set of features and symbols provided by an internal reference or, possibly, via external audio references assigned as such by the listeners (Rumsey 2002). Such references can be the manifestation of top-down sets of prior knowledge (i.e. individual) features. Ideally, mainly due to the short-term duration of auditory memory employed for comparison, any valid judgment of auditory quality has the tendency to be improved by means of direct comparison to a reference template, that is, *benchmarking*. A more detailed exposition of the memory resources employed for internal references (sensory, working, and long-term memory) is provided in Raake and Wierstorf (2014), this volume.

Fig. 3 Stereo image localization accuracy maps derived from a “virtual-listener” binaural perceptual model and simulated reproduction channel data corresponding to **a** the ideal case of omnidirectional loudspeakers under anechoic conditions, **b** 2-way loudspeakers inside an ideal shoebox-shaped control room, **c** the same loudspeakers inside a slightly asymmetric control room of similar dimensions as the one in case **b**. The **dark blue** area indicates highly accurate localization while the **yellow areas** indicate large errors. Each map area corresponds to a $2\text{ m} \times 2\text{ m}$ grid, covering all possible listener positions. The two stereo loudspeakers are depicted in **red** (Kamaris and Mourjopoulos 2018)



However, for practical reasons, this is rarely the case during less formal listening scenarios as are usually performed according to the paradigm of Fig. 1. In most cases, the listeners will then resort to abstract and possibly unreliable references formed from past experience. Additional biasing in the reference set may be due to multimodal stimuli, emotions, and other reference-moderating factors (Rumsey 2002; Blauert 2013). Depending on the application, the listener may judge an individual sound object or multiple sound objects (auditory scenes) at different layers of detail and different layers of internal references. Such multilayer processes as occurring in quality- and/or aesthetic-judgment formation have been described according to the amount of intellectual abstraction involved (Blauert 2013). In ascending order of abstraction, these layers may reflect

1. *Auditive quality*, e.g., assessing features such as loudness, timber (Layer 1)
2. *Aural-scene quality*, e.g., assessing aural-scene transparency, object layout (Layer 2)
3. *Acoustic quality*, e.g., assessing audio and acoustic system response measurements (Layer 3)
4. *Aural-communication quality*, e.g., assessing immersion, functionality and meaning (Layer 4).

This framework particularly applies to judgments made by experienced listeners, such as musicians or audio-mixing and mastering engineers, during the encoding and/or decoding stages of Fig. 1. At the lower layers (Layers 1 and 2), such listeners are able to identify and judge primitive features in sound objects and scenes and can isolate them perceptually with an often remarkable accuracy. As was described in Mourjopoulos (2014), this fact may defy known theoretical laws. An actively performing musician, a composer, or even an audiophile with genuine “golden ears” may additionally utilize physical data for specific encoding/decoding procedures (Layer 3) and also audition at the highest abstraction layer (Layer 4), such as interrelations between acoustic (physical) features of sounds and semiotic or aesthetic features of the content (music), hence further judging the quality and/or aesthetics of sound object or scenes with reference to more abstract cognitive percepts (internal references).

The above approach for a layered framework for audio quality may also accommodate the ambiguous case of audiophile listeners, who often claim to perceive qualitative and/or aesthetic aspects of audio systems and components that may or may not correspond to instrumentally (“objective”) measurable features, hence bypassing some of the abstraction layers representing a manifestation of cognitive indeterminacy in audio technology—compare also Sect. 2.4.

2 Review of Past Work

2.1 *Modeling the Aesthetic Concept and the Observers' Responses*

Aesthetic as a word is stemming from the ancient Greek *aisthēto* (something perceived by the senses) and since the 18th century is the branch of philosophy studying beauty and taste, a branch being closely related to the philosophy of art, which is concerned with the question of how individual works of art are interpreted and evaluated (Munro and Scruton 2018). Traditionally, this universal human trait is defined as *the subjective experience elicited by beautiful stimuli*. This experience can be evaluated by observers on an individual (subjective) scale from “ugly” to “beautiful”. It can be further compared to other stimuli and hence classified on a preference scale (Redies 2015). In everyday life, aesthetics cover a wide range of experiences (visual, literary, musical, auditory, etc.) and perceptual phenomena (natural phenomena, functional objects, aesthetic artifacts, works of art, etc.) (Consoli 2012). The aesthetic experiences mostly relate to human reactions to non-instrumental qualities of an event and address internal processes, multi-sensory properties, psychological aspects, the sociocultural characteristics of its creator as well as of the observer.²

The foundations of experimental aesthetics can be traced to Gustav Theodor Fechner (1876) who, with his book *Vorschule der Aesthetik* (Introduction to Aesthetics), introduced methods relating objective stimuli properties and the aesthetic response (Wikipedia 2018; Graf and Landwehr 2015). Furthermore, since experiences in art and judgments of beauty involve cognition, the cognitive sciences have been also employed for analyzing aesthetic experiences (Stokes 2009; Consoli 2012). Via such analyses, the aesthetic experience is now considered an organizational adaptation and an activity that involves perceptual, cognitive, imaginative, affective and emotional processes. It is based on a specific mental attitude and attentive state, which activated prior to the aesthetic experiences. The attentive state remains active, and supports the information processing dedicated to understanding and interpreting aesthetic objects (Consoli 2012). More recently, contemporary experimental aesthetics have been expanded by research tools, data, and theories from diverse scientific fields such as neurophysiology, visual and auditory perception, psychology, social science, art, brain imaging, semantics, and product design (Redies 2015). Due to such developments, functional models have emerged during the past decade that predict aesthetic responses, emotions, and judgments that elicited by specified stimuli (Leder et al. 2004). Typically, such models incorporate stages that accommodate the previously-mentioned perceptual, emotional, knowledge and cognitive stages. Hence, in contrast to earlier theories of aesthetic preference that were mostly derived from psychology and related fields, current models of aesthetic evaluation consider aesthetic response as the combined effect of sensory processing and internal emotional response.

²Note: There exists an extensive body of literature in the philosophy of aesthetics.

Recently, such models have led to the concept of *processing-fluency* in aesthetic appreciation (Reber et al. 1998, 2004). The aspects of processing fluency are now widely accepted as a basis of mental concepts in experimental aesthetics. This holds in particular for the aesthetics of visual objects with the following propositions.

- Depending on an object's (visual) properties and a beholder's prior experience with this object, the mental processing of the object will be experienced as more or less fluent
- The experience of high processing fluency directly feels good on an affective level
- As long as the positive affect is not attributed to a different source, it infers the aesthetic appreciation of the object, leading the observer to "like" the object in terms of aesthetics (Reber et al. 2004).

The fluency theory has reached popularity in experimental research on aesthetics. This is due to the fact that it allows for models that render predictions of aesthetic judgment as induced by specific stimulus characteristics. It has also become evident from past research that the human aesthetic functionality has evolved as a cognitive adaptation process causing brain functions to be attracted, become attentive to, process more efficiently—in other words, "fluently"—certain evolutionary beneficial stimuli that are associated with pleasure, that is, support a hedonic state-of-mind (Reber et al. 2004; Conrad 2010). Stimuli are classified as beautiful when they trigger and amplify such attentive processes, provided that they conform to specific semantic and contextual principles, either derived from evolutionary biological mechanisms or from top-down cognitive processes. This new theoretical framework for aesthetics has more recently led to functional aesthetic models. These are largely based on analyses of visual images and related artwork (Redies 2015), but also provide a useful groundwork for respective auditory models.

Models based on the notion of *fluency or efficiency of processing* attempt to explain how processing of stimuli is hedonically marked and, consequently, experienced as aesthetically pleasing (Reber et al. 2004). For example, aesthetically significant visual stimuli are coded more efficiently, that is both more easily and more precisely than non-aesthetical ones (Reber et al. 2004). Furthermore, such fluency theory indicates that efficient coding of stimuli from low-level sensory mechanisms up to a cognitive level, is largely implemented on neurobiological mechanisms leading to the psychological phenomenon of fluent information processing. For example, for images and visual art, the observer processes efficiently all physical object features (e.g., color, shape, texture) as well as its semantic, symbolic and narrative elements prior to deriving an aesthetic judgment (Redies 2015).

It is further suggested that at the biological level, aesthetic responses foster the formation or utilization of neural pathways that support mediation between perception and interpretation. According to this model assumption, brain anatomy is not only determined by genetic, but also by epigenetic mechanisms, thus allowing the brain to respond to unforeseen events and to make optimal sense out of the challenges imposed by the world (Conrad 2010).

In order to look more closely to such functional models for aesthetics, it is necessary to discuss the underlying mental processes that mediate between stimuli per-

ception and their interpretative outcome. Clearly, all perception requires transformations, predominantly non-linear ones, that select useful features from the sensory data, filter out unwanted information, monitor the fluency of internal processes, make comparisons on the cognitive level, and so on. Such procedures are to a large extent predictive and top-down controlled (hence largely individually variable) and learned via exposure to the environment (hence culturally variable) (Conrad 2010).

Recent models for visual aesthetic experience combine the above formalistic and contextual aspects of functional aesthetics into a Dual-Process model (Graf and Landwehr 2015):

1. Process 1 *Aesthetics of perception* implements a bottom-up universal mechanism dealing with perceptual processing (e.g., based on the intrinsic form and features of an artwork), which can be evaluated as being or not being beautiful activating a beauty-related mechanism.
2. Process 2 *Aesthetics of cognition* implements a partially top-down mechanism that is variable between individuals according to their cultural experience and is based on cognitive processing of contextual information, such as depicted content, the intentions of the artist and/or the circumstances of the presentation of the artwork.

The processing in the two modes is usually considered as sequential, although each of the processes can be applied independently. The combined outcome of the two perceptual processing stages, *aesthetics of perception* and of *aesthetics of cognition* leads to an optimal resonance and maximum aesthetic appreciation by individual observers as well as among a social group.

Considering such principles for the domain of aesthetic appreciation of sound and music, it becomes immediately evident that the artists' intents, aesthetic choices, and methods provide an initial differentiation between natural sounds and engineered (formalized) acoustic stimuli. Of course, natural sounds can directly be associated with objects, events, scenes, and acoustic sources, but this procedure is not commonly applied for music and its electro-acoustically-generated representation in auditory scenes. Furthermore, there is the significant difference due to the dynamic time-evolution and sequencing of music in contrast to the appreciation of static visual images, which in most cases can be directly associated with natural and well-defined forms appearing in the environment (McDermott 2012). Hence, music semantics and symbols are temporarily evolving, abstract, non-representational and non-depictive. Nevertheless, the effects of music regarding emotional responses and aesthetic appreciation are by no means less substantial than the response to visual-art objects (Eerola 2014).

Note that for any such analysis there is always the danger that the results may be dominated by the aesthetics of musical content and context, composition, form, performance, style, etc. This poses a well-known issue on the research in musicology. However, the aim of the current chapter is to examine how audio-engineering practices, systems and communication channels affect mechanisms that are associated with the aesthetic experience when listening to recorded/reproduced music signals. Hence, care is taken here to separate as much as possible the content (music) from the carrier (audio), and thus de-contextualize the proposed aesthetic models.

2.2 Annotation of the Emotional Responses to Sounds

To be sure, both music in its natural, live-performed form (e.g., when listening to acoustic instruments and voices within real spaces) and their recorded/reproduced representations, share comparable emotionally affecting and pleasure-inducing effects in the listeners (Eerola 2014). As was discussed in the previous section, such emotional responses constitute an important component of the aesthetic experience. In fact, our understanding of fluency and valence in association with functional modeling of the aesthetic experience benefits from current approaches of annotating, evaluating and modeling emotional responses to specified stimuli.

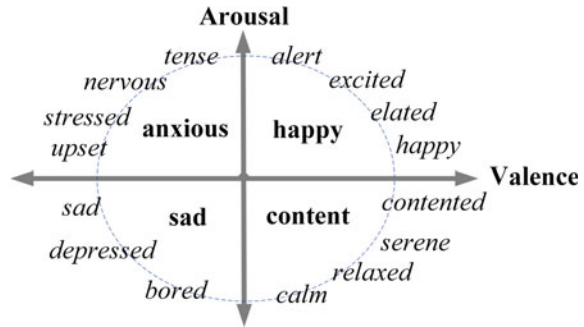
In the context of such functional models (Russel 1980; Posner et al. 2005; Bradley and Lang 2007), the observers' affective states can be defined in terms of a 2-dimensional affective model, the results of which can be plotted in a *Arousal versus Valence* (i.e. interest vs. enjoyment) plane. Such a model is suitable for annotating emotional response to sounds and music (Rumsey 2015a). In relevant work related to spatial-sound reproduction by Drossos et al. (2015), the annotations were performed using the *Self-Assessment-Manikin* (SAM) method, which is suitable for affective-state depiction. Such a simplified but functional approach to clustering and mapping of emotional states is typically based on verbal descriptors as depicted in Fig. 4 (Russel 1980; Posner et al. 2005; Bradley and Lang 2007). In accordance with this model, the existence of the *International Affective Digitized Sounds* (IADS) pre-annotated dataset is relevant—see Bradley and Lang (2007). Emotions and attention labels are contained in this database. They can, for instance, be employed for evaluating observer decisions that can be broadly classified as “similar”. Note that in Williams (2016) a comparable 2D-annotation approach was employed via a three-stage process, namely, for learning, generating, and transforming new musical material as created by analysis of seed pieces of pre-composed music (Rumsey 2015a).

2.3 Quality and Aesthetics of Sounds

A first consideration shows that there are similarities between the concepts of *Sound Quality* and *Sound Character*,³ in the context of aesthetic-appreciation assessment of reproduced audio information. However, it is self-evident for sound and especially for music listening, that aesthetic judgment may be initiated irrespective of any sound-quality judgment and also that judgment on sound quality can be generated independently or complimentary to the assessment of the sound character (Zacharov et al. 2016; Pedersen and Zacharov 2015). Blauert and Jekosch (2012) provide the descriptive definition of sound character as “the totality of measured values of features that are associated with the sound sample under examination”. As was also discussed earlier, these different aspects may include categories such as the

³For details refer, for example, to Raake and Wierstorf (2020), this volume.

Fig. 4 Arousal/Valence (VA) space for emotional annotation clustering (adapted from Bradley and Lang 2007; Drossos et al. 2015)



acoustic and auditory profile, emotional and semantic features and cross-modal cues (Blauert 2013; Blauert and Jekosch 2012; Jekosch 2005).

Although measurable features as the acoustic and auditory profile need assessment by experts, the features generating valence and arousal which are also connected to aesthetic assessment even for non-expert listeners can also function independently to sound-character assessment and are not usually exposed to formal measurement. To a large extent the discriminating ability and response of all listeners to sound and music also relates to aesthetic appreciation, since aesthetics describe human reaction to the *non-instrumental* qualities of a sound event, and the aesthetic experience addresses the internal processes, the multi-sensory properties, the psychological aspects and the sociocultural characteristics of its creator as well as the observer. In contrast, the sound-quality assessment, especially from an engineering perspective, is strongly rooted on objective, instrumental measures usually with respect to predefined references. Furthermore, as noted in Blauert and Jekosch (2012), sound-quality judgment from an engineering perspective can be formed in the context of the suitability of the measured sound features to meet recognized and expected values, i.e. product sound quality is assessed with respect to reference of the “sound of quality”.

2.4 Sound-Quality Assessment

A detailed description of the methods for sound-quality assessment is provided in Raake and Wierstorf (2020), this volume. One of the most critical aspects of established sound-quality-assessment methods relate to the evolution of *sensory descriptors*, and *profiles* of listeners for the development of structured vocabularies (lexica) that can be used to correspond to the sensory evaluation of the specific sound characteristic. Pedersen and Zacharov (2015), Zacharov et al. (2016) provide an overview of such earlier attempts also described in Bech and Zacharov (2006).

As is described in the above references, in building on such lexicon and indirect sensory profiling methods such as *multidimensional scaling* (MDS), the assessors provide scaled evaluation of pairs of stimuli without the necessity for an explicit

definition of the technical or perceptual attributes. Direct sensory profiling methods require exact definitions of attributes, either via consensus vocabulary or individual vocabulary. A number of individual vocabulary techniques exist, such as Free Choice Profiling (FCP), Flash Profiling (FP) and individual-vocabulary profiling (IVP). Pedersen and Zacharov describe the methodology for developing a consensus-based lexicon (termed *Sound Wheel* for reproduced sound) from hierarchical cluster analysis in the semantic space of sound, further validated via statistical correlations during listening tests (Pedersen and Zacharov 2015). The method for selecting such attributes and terms is also conforming to a recent ITU-R recommendation (ITU-R 2017) and significantly, in such lexicon, all pleasure-related (hedonic) descriptors have been removed. Such a measure presents clearly an important differentiator between sound-quality descriptors and aesthetic descriptors that, as was shown, are primarily concerned with assessing hedonic response to sound events. Alternatively to such sensory descriptor-based methods, *explicit-reference-stimuli methods* are used in quality tests implemented by listeners. Methods of this type are the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor, (Bech and Zacharov 2006; ITU-R 2015) and Continuous-Quality Scales (CQS), which can address intermediate quality differences. For sound-quality tests addressing a wide range of quality levels, single-stimulus methods such as the 5- or 9-point absolute category rating (ACR) tests are typically used (Bech and Zacharov 2006). Here, specific stimuli are often presented as hidden references that are not identified as such by the test participants.

In Zacharov et al. (2017) references are provided for additional hedonic scales that can be employed for sensory evaluation especially in conjunction with CQS such as MUSHRA (ITU-R 2015). This study compared a range of different hedonic and quality scales in several experiments for an audio codec with or without a declared reference and audio anchors. The study confirmed a similarity between such different methods (i.e. the labeled hedonic scales, LHS and CQS), but the authors acknowledge that further work will be required especially for the impact of the declaration of explicit reference in the respective scales. Furthermore, they stress the potential for a future hedonic scale “spanning every possible hedonic auditory sensation for the evaluation of all available auditory stimuli”.

Given the well-known practical difficulty in the implementation of tests with listeners, it is also possible to substitute listeners by algorithms and models of perceptual mechanisms and in such cases, the so-called *instrumental methods* are used instead. Such methods implement specific parts of the auditory signal processing, possibly even including some cognition-type mapping to quality dimensions or overall quality. Different elaborate approaches of this type have been developed in the past years, and have been standardized in bodies such as the International Telecommunication Union (ITU). Examples include PESQ and POLQA for speech transmission systems, and PEAQ for audio-coding evaluation (Thiede et al. 2000; ITU-T 2001, 2011; Beerends et al. 2013). These so-called signal-based, full-reference models estimate quality comparing the processed audio signal with an unprocessed reference.

There are many limitations in such quality models (see Bech and Zacharov 2006; Raake and Wierstorf 2014 for more details) and more elaborate perceptual features need to be incorporated along with active exploration and enhanced by top-down

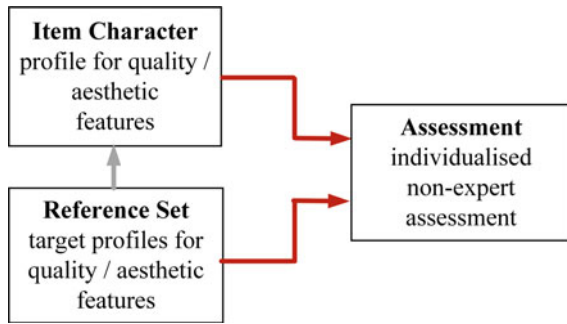
feedback (Blauert et al. 2010; Blauert and Brown 2020). Such recent elaborate models for sound quality utilise *computational models of binaural hearing and perception*, suitable for engineering applications (Raake and Wierstorf 2014; Wierstorf et al. 2013b). For such model, the adoption of top-down, hypothesis-driven functionality can be viewed as enabling the interaction between *black-box* cognitive functions of experience, memory, and adaptation with the sensory binaural data derived from established bottom-up perceptual attributes. These black-box cognitive functionalities are trained via supervised or unsupervised procedures usually based on transformations from signal to symbolic features and utilizing established machine learning and AI classification methods. Such architecture of binaural audition may be able to accommodate the important stages of perceptual inference, knowledge and quality assessment of events, scenes and sounds containing a “brain”, that is, expert components which “interpret” the output of the lower, signal-driven sections of the model. At some cases, this signal-driven (bottom-up) and hypothesis driven (top-down) processing can proceed in an interleaved manner focusing on states which make sense in a given specific situation. Additional top-down feedback paths describe human mental processing such as attention and scene exploration. For a description and a diagram refer to Raake and Wierstorf (2020), this volume.⁴

Hence, such an approach as largely founded on the general architecture for quality assignment, introduced by Blauert and Jekosch (2012) and modified later by Raake and Blauert (2013), is also suitable for adaptation to specific target groups and task-specific applications. This architecture, shown in Fig. 5, accommodates the combined bottom-up and top-down abstraction discussed previously and is suitable for audio/acoustic engineering applications by allowing input from expert measurements/reference sets. In this figure, the “Item Character” box refers to the feature profile of the aural scene, analyzed and assessed via an appropriate metric. The “Reference Set” box refers to the feature profile of the listeners’ expectations and prior knowledge. The “Assessment” box illustrates the extent to which the expectations have been met by the specific item (aural object or scene). As is also discussed in Blauert and Jekosch (2012), Raake and Blauert (2013) this framework can accommodate adaptation of such computational models of binaural hearing and perception to personalized user profiles—further referred to as *model individualization*.

Wierstorf et al. (2013a) discuss the results of two pilot experiments that examined the usefulness of such model with two test cases related to qualitative aspects of spatial-audio reproduction. In a paired-comparison preference test, a musical piece with three sound sources (two guitars and a singer), reproduced over different audio-reproduction systems, was assessed. Thereby, it was shown that the original source scene, as normally used for reference in typical sound-quality tests, was not the preferred choice. Instead, certain degradations of specific sources in the scene were found to be preferred by the test listeners. This study highlights the significance of scene-specific paradigms for sound-quality evaluation and the need for test methods that enable holistic sound-quality evaluations.

⁴Further resources may be found in the project TWO!EARS, www.twoears.eu, which addressed binaural interactive exploration of auditory scenes.

Fig. 5 The “product sound” quality-assignment process. Figure adapted courtesy of Blauert and Jekosch (2012)



The current section indicates that many aspects of sound-quality assessment require enhancement by modeling functions which relate to the aesthetic appreciation of individual listeners.

2.5 Audiophile Aesthetics and Plausibility

A unique group of listeners—the “audiophiles”—usually adopt a “cost-no-issue” approach for accessing “*high-fidelity*”-reproduced sound, thereby not always following properly defined descriptors or judging criteria. This not well-explored phenomenon represents a most extreme example of where the aesthetic and sound-quality assessment of the technological medium takes precedence over the aesthetic appreciation of the content (music) (Hales 2017). Although one would assume, at a first glance, that the ultimate aim of such a dedicated and costly approach to reproduction fidelity would directly correspond to sound-reproduction qualities, namely, utmost authenticity of the reproduced sound with regard to the source material, more careful consideration shows that this is only partially true. In fact, other aesthetic rules are applied by audiophiles. As was briefly discussed in earlier sections, for the case of recordings of “real” instruments and acoustic-scene performances, the authenticity (transparency) requirement must also accommodate the results of technical imperfection and casualness in capturing the event, along with potential mistakes and emotionally-driven excesses from the creators and/or the performers. However, these issues are rarely retained on commercial recordings. In respective commercial products, authenticity and transparency are overruled by technical manipulation and “corrections” of the generic aesthetic quality by engineering practice. Beyond recordings of natural events in real acoustic spaces, synthetic digital sound sources are likely to be involved in recorded music these day. This means, that there is no original auditory event/scene to start with. Instead via the recording and subsequent signal processing an artificial reality is imposed by the creator-engineer, so the final result becomes plausible to him/her—and potentially to the listeners aimed at. In this way, semi-empirical technical concepts regarding the conception of auditory scenes

lead to the formation of an aesthetics of “hyperreal” scenes (Rumsey 2002, 2008; Hales 2017).

With the reference to the multi-layered framework, as introduced above with regard to the assessment of audio quality, respectively, the case of audiophile listeners may or may not correspond to instrumentally (“objective”) features in the lower layer of the multilayer scale, but instead, the assessment is based on the aural-scene quality layer (e.g., aural scene transparency, object layout, plausibility). As is also discussed in Raake and Wierstorf (2020), this volume, sets of reference features are evoked by the listeners’ expectations in a given listening context, and are related to perceived features which are not always “nameable”. Bearing this in mind, it is necessary to reconsider the principal differences between the cases where the listeners have access to a “ground-truth” reference for direct comparisons of quality/aesthetics of any simulated auditory scene. In the case that such reference is not available, the aspect *plausibility* of the presentation is usually invoked by the listeners as a substitute.

At a philosophical level encompassing all fields from science to art, these differences are discussed in Dutton (1977). Yet, as explained in detail in Lindau and Weinzierl (2011), a system-oriented criterion would differentiate the concepts of authenticity and plausibility (of a virtual environment). Thus for such applications, plausibility is defined as a simulation in agreement with the listeners’ expectation towards a referential “real” acoustic object or scene. Such internal references relate to each listener’s personal experience and expectations rather than to exact perceptual identities. In many cases this is taken as sufficient for evaluating the quality of a simulation. As will be discussed in Sect. 3.2, experienced observers following lengthy exposure to specific stimuli, develop a level of *cognitive mastering* (Leder et al. 2004). This process is also referred to as *assimilation*, representing the fitting of the perceptual representation to existing conceptual mental patterns (*schemata*) (Sotujo et al. 2020). With respect to the layered framework for audio quality/aesthetic assessment discussed earlier, Raake discusses different levels of build-up mechanisms (Raake and Wierstorf 2014), this volume. These can develop via

- *Passive indirect build-up* via exposure to different systems
- *Action-selection build-up* via system comparison
- *Active-control build-up* via modification of systems and events

Thus even when no direct external reference is provided for comparison, any criticism or appraisal such as applied by experienced listeners, creators, or audiophiles, largely relies on internal references accessed largely via cognitive “top-down” processes. These are definitely involved in aesthetic judgment, since all these observers invoke and project their personal experiences, expectations, plausibility precepts, and affective interpretations into internal references. Limitations with respect to the ground-truth criterion for audio aesthetics are discussed in Francombe et al. (2015), where an elicitation experiment was performed to determine the qualitative differences between the experience of listening of real versus reproduced audio. The results “... highlighted the many differences between real and reproduced audio in terms of

timbral, spatial, and other factors ... For methodological reasons, it is difficult to make direct comparisons between real and reproduced audio ...”. This outcome was further elaborated with respect to the spatial-reproduction mode in Francombe et al. (2017) where it is noted that such differences between reproduction methods may be related to the level of listener experience. Further, for the experienced listeners perceptual attributes that are associated with the presentation technology, such as *listener envelopment*, turned out to be significant.

2.6 *Spatial-Sound Aesthetics—Envelopment, Immersion and Emotional Inhibition*

As was already discussed in Sect. 1.3, recent encoding formats beyond stereo offer an enhanced auditory representation of spatial source and scene properties. Thus formats like discrete multichannel, WFS, HOA, binaural with headphones) offer increased listening envelopment, and such the sense of immersion (Rumsey 2015b, 2016, 2017). At first consideration, envelopment, immersion by rendering virtual sources in three dimensions (3D), in particular beyond the frontal-azimuth angles, increases arousal and valence (Drossos et al. 2015) hence contributing additional factors in the aesthetic experience. However, due to strong perceptual precedence generated by sounds delivered at such angles employed by audio engineers, principal or interrupting sonic events are not allocated at such periphery channels, except if an intentionally extreme emotional response is required—for instance, for creating special audiovisual, cinematic plots for “home theater”. Clearly, arousal is a vital aspect of aesthetic experience both at the perceptual (“head turning”) level and at the cognitive level of judging on such events. By the way, it is well known that arousal can be also initiated at the sensory level, especially when spatial signal cues are artificially modified, for instance, by applying artificial reverberation, delays, or extreme stereo panning.

To validate such effects, tests reported in Drossos et al. (2012, 2014, 2015) included a series of listening experiments using samples that form an emotionally labeled sound-data base with 167 sounds having multiple semantic contents (Bradley and Lang 2007). Binaural processing of the original sound data was realized for five binaural versions, corresponding to azimuths of 0, 45, 90, 135, and 180°. Finally, the participants’ affective state was defined in terms of a 2D model, while the annotations were performed using the *Self-Assessment Manikin* (SAM) method (Russel 1980; Posner et al. 2005), which is suitable for affective-state depiction.

This binaural sound corpus was emotionally annotated through a series of online auditory-evaluation experiments using a custom web platform and headphone presentations, with the participation of listeners of different cultural background. For the emotional rating tests, 215 participants responded with ≈ 3000 valid annotations.

The results for the class of natural sounds revealed that sonic events impose a systematic effect on the listeners’ emotional states when they move towards the

lateral limits of the field of vision. This change is either an increased activation, *arousal*, combined with a lowered pleasantness, *valence*. However, the exact opposite has been also observed, that is, an increased pleasantness combined with a lowered activation. The number of events assigned to either case was found to increase with the angular horizontal position of the sound source. In addition, when the source was located exactly at in the rear of the listeners, the number of auditory events that caused increased arousal and lower valence was greater than the amount of those that increased valence and decreased arousal.

For the specific class of presented music events (audio segments), the results indicated that listeners feel more pleasure when the sound source moves from zero-azimuth presentation towards a side position with a peak in valence at 45°, indicating the emotional advantage of an expanded spatial field in front of the listeners. This effect is also achieved in a typical stereo set-up. However, auditory events beyond 90°, increase valence. This indicates a relative unpleasantness, for instance, for a singing voice panned beyond the visual field. Especially for extreme azimuthal positions such as straight behind at 180°, valence was found to decrease, although arousal was not systematically affected. These preliminary findings correlate with the amplitude-panning practice employed in audio-mixing when utilizing typical multichannel setups—such as, in 5.1 surround sound. It is well known that in audiovisual content mixing for the “home theater” virtual sound events are panned to the rear channels for increased arousal, emotional involvement and decreased valence, whereas music events are mostly spatialized to appear in a frontal angular position. The arousal emotion can be also linked to fear, as was demonstrated in Ekman and Kajastila (2009), where tests were conducted to evaluate the emotional impact of sounds in games.

Similar aspects were also discussed in Lepa et al. (2014). There it was found that when listening to music pieces of different genres and valences in different spatial-presentation modes the live-concert simulation was found to be the most intense with regard to all four dimensions of perceived affective musical expression compared to stereo reproduction over headphones leading to an increase in perceived emotional expression of music—in addition to an increase in perceived spatial quality. Furthermore, quality expectations seemed to leave the expressional dimension of music unaffected, yet, leading to an increase in the attributed quality. Hence, it was concluded that technology-related placebo effects apply for music, namely, with respect of the perceived audio quality but not of the perceived emotional expressivity. Thus, these placebo effects concern increased envelopment and immersion via spatialization and reproduction. In other words, the findings regarding the respective effects were confirmed, namely, “. . . to rely more on the additional phenomenal quality of externalization itself, the playback technology allows . . . the feeling of being part of an auditory scene surrounding one’s own body . . . than on improvement in spatial auditory-scene detail . . .”.

For object-based audio rendering, as has been adopted by the MPEG-H standard (Herre et al. 2015), in Francombe et al. (2018) it is shown that *envelopment* is one of the most important attributes for the listener preference of spatial audio, irrespective of the complexity of the reproduction system used.

Complex interactions between context, presentation and acoustics are described in Herre et al. (2015). Such contextual bias-build-up was found to exist over the course of minutes and, since sound localization is a dynamic process that depends on both the context and the level of reverberation in the environment, interactions between sequential sound sources occur on time scales of hundreds of milliseconds up to minutes—Kopco et al. (2007). A later study showed a complex effect of the temporal characteristics of the context on sound localization, probably driven by processes in multiple stages of the auditory pathway. Such findings provide challenging tasks for models of spatial auditory perception. Usually, current models do not consider processing over a wide range of time scales or multiple forms of adaptation operating at the same time. Additionally to the above considerations, the recent developments in VR and immersive visual-media technologies introduce further open questions with respect to audiovisual attention, cross-modal features and quality interactions—Raake and Wierstorf (2020), this volume.

2.7 Functional Aesthetics in Images, Music and Sound Technology

The previous sections highlighted the importance of evolving models that incorporate bottom-up modeling of peripheral auditory mechanisms along with modeling of top-down cognitive functions in order to implement a realistic framework for sound perception and the appreciation of quality and its aesthetic properties.

Enrichment by the aesthetic functionality may enhance existing sound-quality evaluation methods and allow special kinds of experiences to be implemented for the appreciation of objects of art, sounds, music, and their carrier media. In the previous sections it has been shown that aesthetic appreciation at a functional level can be considered as a complementary dimension in the quality-evaluation process of any natural or man-made object, artifact, or work, both adding up to an integral sum of its properties, but also including aspects of the cognitive functions of the listeners.

In order to construct models of aesthetic judgments, it is paramount to consider evidence stemming from empirical, clinical results as well as from computational analysis of the dominant sensory modalities, namely, vision and hearing.

At first, it is useful to consider the relevant body of research as carried out on visual aesthetics. This research has followed different approaches, ranging from algorithmic modeling of aesthetics for the appreciation of images and pictures (Deng et al. 2017) to EEG studies of brain response to conceptual art (Kontson et al. 2015; Renoult et al. 2016).

An extensive overview of recent computer-vision techniques used in the assessment of visual-image aesthetic qualities is provided in Deng et al. (2017). An aim of this work was to survey experimental results of methods that can derive binary classifications for quality judgment on photos, for example, to distinguish between

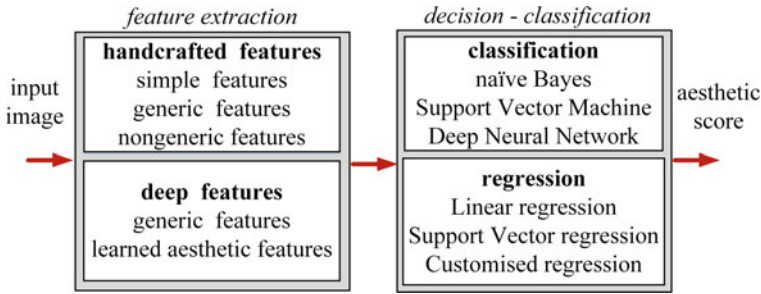


Fig. 6 Typical flow of image-aesthetics assessment systems (adapted from Deng et al. 2017)

high- and low-quality data sets. Computer-vision researchers adopted a classification/regression approach for such assessment, driven by a training stage that relies on image-features extraction, manually or automatically, from high and low-quality image datasets. The performance accuracy is judged on the basis of a metric that assesses the classification outcome from manually or data-driven annotated datasets. As is shown in Fig. 6, different approaches can be adopted for image-feature extraction and representations as well as for the feature-classification stage.

Considering perceptual attributes that can be utilized for the global analysis of images, Renoult et al. (2016) found out that the algorithmically modeled sparseness of the activity of simple cells in the primary visual cortex (V1) correlates with female face attractiveness when assessed by male participants. This suggests that there might be general, non-face-recognition-specific neuronal properties that affect facial aesthetic evaluation. Other candidates for global sensory properties that have been studied recently include processing-fluency distribution of spectral-frequency power, self-similarity, and fractal properties (Redies 2007, 2015; Reber et al. 2004; Renoult et al. 2016; Kiebel and Friston 2001; Rao and Ballard 1999). Further related work can be also found in Moon and Spencer (1944), Nishiyama et al. (2011), Mavridaki and Mezaris (2015).

Moving on to relevant work in to music, a report by Tiihonen et al. (2017) provides results of structured literature-reviews of past empirical studies in diverse scientific fields that are related to pleasurable, hedonic, enjoyable, and rewarding experience of music and the visual arts.

This literature review aimed at the understanding of how pleasure derived from music and visual art had been understood conceptually, either directly or indirectly, in empirical research during the past 20 years. The results indicate that in the visual-art literature, pleasure was employed vaguely, that is, many times it was not a clear object of the investigation but rather a characterization of the researched phenomenon. In contrast, music research conceptualized pleasure by identifying elements of core hedonic response (valence and arousal, see Sect. 2.2) and intrinsic reward. Hence, it has been confirmed that music is able to activate the reward center. This was accomplished by psychophysiological measures referring to the notion that the pleasure that music induces is to a certain degree biologically based, rather than culture and context

specific. This finding is in agreement with other research evidence that indicates that biological mechanisms are associated to hedonic responses to music (Levitin 2011). As has been demonstrated by Martínez-Molina et al. (2016) with tests of autonomic nervous-system activities (e.g., skin-conductance response, heart rate measurements, fMRI scanning and psychometric questionnaires) applied to listeners with a specific “musical anhedonia”, music enjoyment and pleasure is primarily associated with the interplay between the auditory cortex and the subcortical reward network. This is an indication of significant biological processes on the level of inhibition, specific for each individual person.

Some directions and guidance with respect to potential methodological approaches for common analysis and modeling approaches in the two sensory modalities (vision and hearing) were provided by Brattico et al. (2017). In this work, the authors state that studies of visual aesthetics indicate that the experience of visual beauty is grounded on global statistical and computational properties of the stimulus, for example, scale-invariant Fourier spectra or self-similarity of image signals. For visual aesthetics, the main contributing factors are assumed to be global across a visual object (i.e., they concern the percept as a whole), formal or non-conceptual (i.e., concerning form rather than content), computational and/or statistical, and based on relatively low-level sensory properties. The authors suggest that studies of the aesthetic responses to music could benefit from the same approach. Thus, along with local signal features such as pitch, tuning, consonance/dissonance, harmony, timbre, or beat, additional global sonic properties can be seen as contributors to aesthetic musical experience. Such approach calls for global-scale analyses in music aesthetics, in other words, to the notion that the impression is mainly generated by the “whole sound”, as opposed to the assumption that it is based on the evaluation of any of its individual components such as specific instruments, harmony structures, intervals, melodies, tuning, rhythm.

Some recent work has analyzed musical stimuli in terms of their global sensory properties. The investigations are based on experiments during which the participants were required to listen attentively to a whole piece of music while their brain-signal activity was measured—for instance, with fMRI. The brain signals were analyzed as time-series and compared to audio music-signal features obtained through music-information retrieval (MIR) analysis (Alluri et al. 2012). During the tests, six perceptual features, *fullness*, *brightness*, *timbral complexity*, *key clarity*, *pulse clarity*, *activity*, and *dissonance* underwent a *principal-component analysis* (PCA).

3 Modeling Audio-Aesthetics

3.1 A Combined Model for Listener Quality and Aesthetic Assessment

From Sects. 2.3 and 2.4 it becomes evident that there is strong complementarity in sound-quality assessment and the aesthetic appreciation of recorded sounds. Current

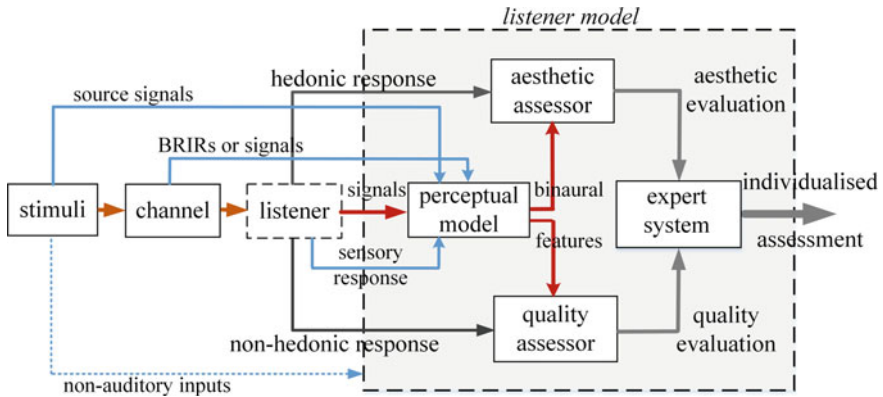


Fig. 7 General conceptual model for quality and aesthetic assessment of recorded and reproduced sound, based on a proposal for parallel assessment for aesthetic and quality

sound- quality-evaluation methods require assessment of objective and non-hedonic parameters of signals, removing personal preferences, affective response, emotions or psychological biasing of the observer. In contrast, aesthetic appreciation relies on personal preferences and emotional response (e.g., arousal and valence), that is, on the hedonic responses of the observers—often formed by social and cultural references. Depending on the specific judgment case, both non-affective/non-hedonic/quality assessment and the affective/hedonic/aesthetic assessment may be applied either separately or in a complementary fashion by the observer. The outcome is thus combinatorial in some way.

Examples of individual aesthetic and quality assessment for the reproduction of the scenario in Figs. 1 and 2 can be recordings that have been re-issued in different spatial/channel formats (e.g., original stereo CD-DA to multichannel SACD formats). Such an example is the recording of Pink Floyd, “The Dark Side Of The Moon, EMI-7243 582136 2 1, and EMI-582 1362”. Comparing such different options, the listeners may apply a sound-quality assessment independently and/or relatively to the source-scene presentation via 2 or 5.1 channels. They may also derive an aesthetic assessment of the spatial merits of each presentation irrespective of any judgment of the sound quality of the recording. In all cases, the technical merits of loudspeaker response, that is, placement, room acoustics, etc., will also affect such assessments and may even introduce additional factors for the assessment.

For any future model to accommodate both such functionalities, it is here suggested to provide two parallel branches, one dealing with sound-quality assessment and the other one with aesthetic appreciation—as is shown in Fig. 7.

Figure 7 illustrates the formation of the individualized aesthetic assessments, modeled in terms of a “conceptual mixer”, providing both aesthetic and qualitative judgments in parallel. For any subsequent modeling of such cognitive mixer for the observer responses, the “blackboard” model employing individualized “experts” can be used (Schymura and Kolossa 2020), this volume. In the left of the diagram, the

signals/stimuli reproduced via the channel (see also Figs. 1 and 2) can be considered as inputs to a human listener. This route is activated only when the human assessment is elicited via *explicit-reference-stimuli methods* and *Continuous Quality Scale (CQS)* methods following non-hedonic and hedonic descriptors—see Sect. 2.4 and Bech and Zacharov (2006), ITU-R (2015, 2017). Such a stage can be bypassed when the listener is substituted by a computational model. In such a case, the diagram can accommodate the existing *instrumental* sound-quality-assessment methods such as PEAQ—Sect. 2.4 and Thiede et al. (2000), ITU-T (2001, 2011), Beerends et al. (2013) via model comparisons between source-input and channel-output signals. When *computational models of binaural hearing and perception* are employed—see Sect. 2.4 and Blauert et al. (2010), Raake and Blauert (2013), Wierstorf et al. (2013b), the channel BRIRs and the source signals can drive the relevant model stage. In all cases, the proposed approach for the listener model is subsequently split into two parallel streams, one for the hedonic and the other one for the non-hedonic features, driving assessors to provide complementary evaluations. The results are finally combined in an expert-system module which renders the individual aesthetic-quality assessment of a specific listener. This module can incorporate knowledge sources, for instance, in a blackboard system or any other kind of a decision/classification model—compare Fig. 6.

As depicted in Fig. 7, the listener may respond with the methods as discussed in the previous section, namely, via psychometric/lexicon procedures driven from the audio data (using original sources as reference, if required), and/or by analysis of such data by a “virtual listener”, that is, a computer model of binaural hearing and perception (Blauert et al. 2010, 2013; Raake and Blauert 2013). An open question relates to the degree and functional description of the “cognitive-mixing” function of the observer, in particular, regarding the degree and the way in which this model stage utilizes and combines the output of each of the two parallel branches. As was discussed above, such a “mixing” function may be controlled by internal plausibility references of listeners. Additionally, the listeners’ individualized aspects of stimulus character and suitability may be accommodated via the “product-quality” procedure described by Fig. 5. A more complete model of individual observers may even incorporate personalization in terms of an individual balance of “objective” and “subjective” judgments.

3.2 *Modeling the Aesthetic Responses to Audio-Scene Presentations*

For further analysis of the “aesthetic assessor” model, it is proposed to utilize a *dual-process approach* after Graf and Landwehr (2015) with two distinct assessment processes, namely, one for *the assessment of the aesthetics of in terms of perception* and another one for *the assessment of the aesthetics in terms of cognition*.

Such a preliminary model is in accordance with the distinction for the mental processes adopted by the Dual-Process model for aesthetic appreciation as was presented in Sect. 2 and is shown in Fig. 8. This technically functional model is driven by binaural signal features and parameters extracted from the *perceptual-model module* utilizing current methods for modeling the periphery of the auditory system. From internal representation parameters provided by these periphery-model stages, the *sensory-assessor* block, as depicted in Fig. 8, will extract metrics of sensory fluency. As it is suggested in Graf and Landwehr (2015), the sensory assessor accounts for the processing that is implemented automatically upon receiving the stimulus and is occurring without the perceiver's intention to do so, and without requiring the perceiver to invest considerable amounts of cognitive capacity or memory resources (Zajonc 1980). An extreme case for activation of this stage during listening to spatially reproduced sound, for instance, is the “head turning” reflex, which helps to avoid front/back confusion in sound-source localization—compare Blauert and Brown (2020).

When a stimulus receives sufficient attention from a listener, the second stage, that is, the stage of top-down-controlled processing, may subsequently be activated and potentially overwrites the automatic responses of the previous stage. Figure 8 illustrates this with the *fluency-assessor* block. This block involves higher-order cognitive processing associated with a detailed and deliberate stimulus analysis and the assignment of meaning. This requires a high amount of cognitive capacity and demands working-memory resources. This type of processing is associated with active and reflective interaction, acquaintance with the stimulus, and potential adaptation or updating of the observer's cognitive structures. This leads to what Leder et al. (2004) have termed “cognitive mastering”. This state may be attained by expert listeners after lengthy exposure, for instance, by musicians, audio engineers, and audiophiles. Overall, such processing translates into feelings of fluency due to an internal monitoring system that screens, integrates and summarizes the ongoing difficulty or ease that goes along with the processing of aesthetic stimuli with respect to the references of the individual listener references.

Hence, whereas *perceptual fluency* relates primarily to the physical identity of the stimuli, *conceptual fluency* relates primarily to assigning meanings to stimuli and to correlate these with the listeners' knowledge-based references on a semantic level. It has been proposed by Graf and Landwehr (2015) that, depending on the fluency level, the outcome of perceptual aesthetic processing can be simplified as a binary positive (pleasure) or negative (displeasure) decision, related with the emotional annotation procedures described earlier in this chapter. Depending on the outcome of this first stage and depending on the specific stimuli, the listeners will or won't be motivated to initiate cognitive evaluation processes as the second stage of their aesthetic assessment.

To be sure, the preliminary model architecture in Fig. 8 presents a highly simplified structure. It is also likely that the observers' motivation to process stimuli in a top-down-controlled way is determined by the interplay of the perceivers' need for cognitive enrichment and the fluency-based affective response to the stimuli. Potentially, disfluency may be reduced by a need for cognitive enrichment and expectations—

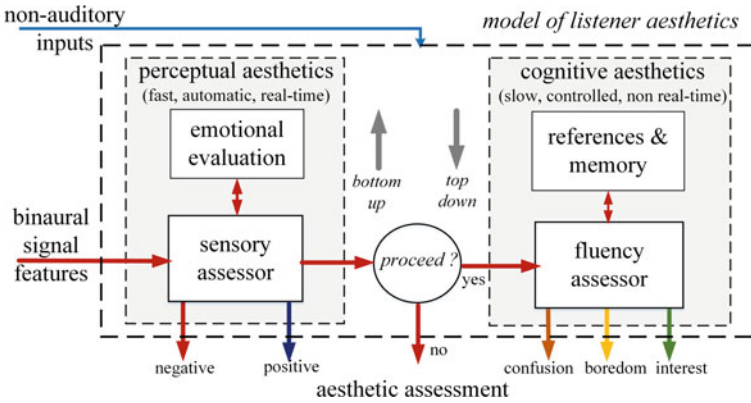


Fig. 8 Proposed functional model for assessment of the aural aesthetic experience. Adapted from Graf and Landwehr (2015)

see Graf and Landwehr (2015). Hence, during both processing stages, the fluency expectations of the listeners are compared to fluency properties of the stimuli.

Any outcome of frustration due to a failing of making progress in processing the stimuli, that is, if no fluency differential appears although the perceiver was motivated to invest cognitive effort, then this leads to a hedonically-negative connotation of confusion, which is experienced as an aversive result. Like with confusion, the nonexistence of a fluency differential during controlled processing can be related to boredom, although in contrast to confusion, in the case of boredom due to a lack of time-variability, fluency occurs at a constantly high level. This combination also leads to a hedonically-negative connotation.

Although a detailed implementation of a model of aesthetic evaluation and assessment is a topic of future research, from the above it is evident that the output of such a model can so far only provide limited—even binary—decisions. In contrast to the assessment of static images, it is also clear that such output may be temporarily varying due to the time-varying nature of audio signals and the resulting dynamically varying cues. However, it is still unclear, how such individual assessments are capable of providing an overall aesthetic rating, for example, of a complete piece of music. Nevertheless, current methods for assessing the rate of change and the adaptation of feature vectors, along with machine-learning techniques, offer options for advanced algorithms to be included in such a model.

4 Summary and Concluding Remarks

Since the beginning of the 20th century, the audio-reproduction technology has achieved wide acceptance. This is not least due to the generated impact, plausibility, realism and arousal that is matching and often exceeding the level of emotional

involvement as achieved via natural listening to sounds and, in particular, to music. Clearly, music, both in natural, live performed form, and its recorded-reproduced presentation, shares comparable potential for emotionally affecting and inducing pleasure to the listeners. Such an emotional response constitutes an important component of the aesthetic experience.

With the emerging formats for holophonic and multichannel reproduction, auditory-scene envelopment and rendering of virtual sources beyond the frontal azimuth angles increases arousal and valence and, hence, contribute additional factors to the aesthetic experience. With traditional stereo reproduction, arousal can be also enhanced at the sensory level, especially when spatial signal cues are artificially modified, such as by artificial reverberation, delays, and extreme stereo-source panning. However, it was also found that binaural rendering can generate more intense experiences compared to stereo reproduction, and that technology-related placebo effects apply mainly to the perceived audio quality and less to the perceived emotional expressivity of music. It is thus becoming clear that many aspects of sound-quality assessment require additional components beyond classification and grading by means of objective metrics. Consequently, this holds for the listeners as well.

From the literature in diverse scientific fields related to aesthetic appreciation, it can be concluded that the human aesthetic functionality has evolved as a cognitive adaptation process to allow brain functionalities to be attracted and process efficiently. Fluency turns out to be a crucial factor in this regard. Certain evolutionary beneficial stimuli can be also associated with pleasure—often referred to as hedonic stimuli. Stimuli are classified as beautiful when they trigger and amplify such attentive processes, provided that they conform to specific semantic and contextual principles—either derived from evolutionary biological mechanisms or from top-down cognitive processes. Current research indicates that biological mechanisms are associated with hedonic response to music, as was demonstrated via tests on autonomic nervous-system activity, for example, by employing skin-conductance responses and heart-rate data, along with fMRI scans and psychometric questionnaires. These tests indicate that music enjoyment and pleasure is primarily associated with an interplay between the auditory cortex and the subcortical reward network, and is formed by biological mechanisms that control the level of inhibition in each individual person.

Recent functional models that have evolved to describe how processing of stimuli is hedonically marked and experienced as aesthetically pleasing are based on measures of the rate and the efficiency in coding and classifying the observers' emotional states.

Hence, potentially measurable aspects of adaptation and accuracy in fluency and valence can be employed in functional modeling of aesthetic experience. Note that, as was also described above, emotional response can be incorporated within such functional models by defining the observers' affective states in terms of a simplified 2D valence/arousal space. In highly simplified terms, the proposed fluency codec provides decisions fluency—*interest* (valence, arousal)—or for non-fluency by processing the specific stimuli derived from their internal representation. Positive affect results only if progress occurs at an extent that is higher than the standard. The opposite, leads to negative affect. When the progress occurs at a rate that was expected, for

instance, as compared to predetermined thresholds, no affective reaction is elicited. Any fluency discrepancy with respect to expectations provides the individual cues that inform the listener about their affective feeling. Failing to make progress in processing the stimuli, irrespective of the observers' motivation to invest cognitive effort, leads to a hedonically negative connotation of confusion, namely, aversion. Like confusion, high constant fluency with no variation indicates boredom and, hence, generating hedonically negative connotation.

Driven by the notion of fluency, the literature proposes that functional aesthetic modeling can be separated into two processing stages. The first one provides for perceptual-aesthetic coding, being activated in real-time and autonomously by a bottom-up response to the sensory features of the stimuli. The second one provides for cognitive aesthetic coding. It comes into play after deliberate non-real-time analysis in a continuous, fluency-controlled adaptive fashion, mostly utilizing top-down inferences, memory, preferences, and expectations. Apparently, the second processing stage is activated when the initial assessment from the first process exceeds specific thresholds of attention.

Although such broad framework for aesthetics has recently led to functional aesthetic models which were largely for the analysis of static visual images and related artwork, it is believed that they can be adapted to the analysis of auditory stimuli—complementary to predictors of sound quality for sound reproduction. Currently, sound-quality assessment methods are strongly rooted in objective, instrumental measures, usually with respect to predefined references that intentionally exclude the observers' emotions, preferences, and hedonic responses.

However, state-of-the-art binaural computational models for auditory scene analysis, as is discussed in detail in Schymura and Kolossa (2020), this volume, address bottom-up perceptual and higher-level top-down cognitive mechanisms, thus providing for a structure that can also accommodate the aforementioned functional model of aesthetic appraisal. Hence this chapter proposes that these computational models could be extended to incorporate aesthetic functionality beyond or in conjunction with qualitative assessment, thus enabling to include human emotional reaction to the inhibition arousal, the pleasure-inducing (hedonic) biological mechanisms, and those qualities of sound events that cannot be measured with instrumental methods. It is likely that the overall listener evaluation of the quality, character and aesthetic connotations of a reproduced sound source or auditory scene will be strongly mediated by a plausibility precept, initiated via top-down individualized “ground truth” references, which are driven by the listeners' past experiences and expectations. Nevertheless, the way in which these two aspects of aesthetics, that is, “objective” and “subjective”, can be combined in computational models, remains a challenging topic for future research.

Acknowledgements The author acknowledges the help of Konstantinos Kaleris, postgraduate researcher at the Audio and Acoustic Technology group, with the preparation of this manuscript. The author also wishes to thank Jens Blauert, Ruhr-University of Bochum and the two anonymous reviewers for useful comments, suggestions and corrections.

References

- Alluri, V., P. Toiviainen, I.P. Jääskeläinen, E. Glerean, M. Sams, and E. Brattico. 2012. Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage* 59 (4): 3677–3689. <https://doi.org/10.1016/j.neuroimage.2011.11.019>.
- Bech, S., and N. Zacharov. 2006. *Perceptual Audio Evaluation: Theory, Method and Application*. New York: Wiley.
- Beerends, J.G., C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl. 2013. Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part II—perceptual model. *Journal of the Audio Engineering Society* 61 (6): 385–402.
- Bertet, S., J. Daniel, and S. Moreau. 2006. 3D sound field recording with higher order Ambisonics – objective measurements and validation of spherical microphones. In *Audio Engineering Society Convention 120*.
- Blauert, J. 2013. Conceptual aspects regarding the qualification of spaces for aural performances. *Acta Acustica United with Acustica* 99 (1): 1–13.
- Blauert, J., and G.J. Brown. 2020. Reflective and reflexive auditory feedback. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 3–31, Cham, Switzerland: Springer and ASA Press.
- Blauert, J., and U. Jekosch. 2012. A layer model for sound quality. *Journal of the Audio Engineering Society* 60.
- Blauert, J., J. Braasch, J. Buchholz, H. Colburn, U. Jekosch, A. Kohlrausch, J. Mourjopoulos, V. Pulkki, and A. Raake. 2010. Aural assessment by means of binaural algorithms – the AABBA project. In *Binaural Processing and Spatial Hearing, Proceedings of the 2nd International Symposium on Auditory and Audiological Research – ISAAR'09*, ed. J. Buchholz, T. Dau, J. Dalsgaard, and T. Poulsen, 113–124. Ballerup, DK: The Danavox Jubilee Foundation.
- Blauert, J., D. Kolossa, K. Obermayer, and K. Antiloglu. 2013. Further challenges and the road ahead. In *The Technology of Binaural Listening*, ed. J. Blauert, 477–501. Berlin, Heidelberg: New York: Springer; ASA Press.
- Bradley, M.M., and P.J. Lang. 2007. *The International Affective Digitized Sounds (2nd Edition: IADS-2): Affective Ratings of Sounds and Instruction Manual*.
- Brandenburg, K., F. Klein, A. Neidhardt, U. Sloma, and S. Werner. 2020. Binaural attention control via congruence/incongruence. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 623–663, Cham, Switzerland: Springer and ASA Press.
- Brattico, E., and M. Pearce. 2013. The neuroaesthetics of music. *Psychology of Aesthetics, Creativity, and the Arts* 7 (1): 48–61. <https://doi.org/10.1037/a0031624>.
- Brattico, E., B. Bogert, and T. Jacobsen. 2013. Toward a neural chronometry for the aesthetic experience of music. *Frontiers in Psychology* 4. <https://doi.org/10.3389/fpsyg.2013.00206>.
- Brattico, P., E. Brattico, and P. Vuust. 2017. Global sensory qualities and aesthetic experience in music. *Frontiers in Neuroscience* 11.
- Breebaart, J., and C. Faller. 2007. *Spatial Audio Processing: MPEG Surround and Other Applications*. New York: Wiley.
- Conrad, D. 2010. A functional model of the aesthetic response. *Contemporary Aesthetics* 8.
- Consoli, G. 2012. A cognitive theory of the aesthetic experience. *Contemporary Aesthetics* 10.
- Daniel, J., S. Moreau, and R. Nicol. 2003. Further investigations of high-order Ambisonics and wavefield synthesis for holophonic sound imaging. In *Audio Engineering Society Convention 114*.
- Deng, Y., C.C. Loy, and X. Tang. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine* 34 (4): 80–106. <https://doi.org/10.1109/msp.2017.2696576>.
- Drossos, K., A. Floros, and N.-G. Kanellopoulos. 2012. Affective acoustic ecology. In *Proceedings of the 7th Audio Mostly Conference on a Conference on Interaction with Sound - AM 12*.

- Drossos, K., A. Floros, and A. Giannakouloupoulos. 2014. Beads: A dataset of binaural emotionally annotated digital sounds. In *IISA 2014, The 5th International Conference on Information, Intelligence, Systems and Applications*.
- Drossos, K., A. Floros, A. Giannakouloupoulos, and N. Kanellopoulos. 2015. Investigating the impact of sound angular position on the listener affective state. *IEEE Transactions on Affective Computing* 6 (1): 27–42. <https://doi.org/10.1109/taffc.2015.2392768>.
- Dutton, D. 1977. Plausibility and aesthetic interpretation. *Canadian Journal of Philosophy* 7 (2): 327–340. <https://doi.org/10.1080/00455091.1977.10717022>.
- Eerola, T. 2014. Modeling emotions in music: Advances in conceptual, contextual and validity issues. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*.
- Ekman, I., and R. Kajastila. 2009. Localization cues affect emotional judgments – results from a user study on scary sound. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*.
- Faller, C. 2004. Parametric coding of spatial audio. Ph.D. thesis, EPFL, CH-Lausanne.
- Francombe, J., T. Brookes, and R. Mason. 2015. Elicitation of the differences between real and reproduced audio. In *Audio Engineering Society Convention 138*.
- Francombe, J., T. Brookes, and R. Mason. 2017. Evaluation of spatial audio reproduction methods (part 1): Elicitation of perceptual differences. *Journal of the Audio Engineering Society* 65 (3): 198–211.
- Francombe, J., T. Brookes, and R. Mason. 2018. Determination and validation of mix parameters for modifying envelopment in object-based audio. *Journal of the Audio Engineering Society* 66 (3): 127–145.
- Graf, L.K.M., and J.R. Landwehr. 2015. A dual-process perspective on fluency-based aesthetics. *Personality and Social Psychology Review* 19 (4): 395–410. <https://doi.org/10.1177/1088868315574978>.
- Grosse, J., and S.V.D. Par. 2015. Perceptually accurate reproduction of recorded sound fields in a reverberant room using spatially distributed loudspeakers. *IEEE Journal of Selected Topics in Signal Processing* 9 (5): 867–880. <https://doi.org/10.1109/jstsp.2015.2402631>.
- Hales, S.D. 2017. Audiophile aesthetics. *American Philosophical Quarterly* 54 (2), 195–206.
- Hamilton, A. 2003. The art of recording and the aesthetics of perfection. *The British Journal of Aesthetics* 43 (4): 345–362. <https://doi.org/10.1093/bjaesthetics/43.4.345>.
- Herre, J., J. Hilpert, A. Kuntz, and J. Plogsties. 2015. MPEG-H audio—the new standard for universal spatial/3D audio coding. *Journal of the Audio Engineering Society* 62 (12): 821–830.
- ITU-T. 2001. Perceptual evaluation of speech quality (PESQ). ITU (International Telecommunication Union), Geneva, Switzerland.
- ITU-T. 2011. Perceptual objective listening quality assessment (POLQA). ITU (International Telecommunication Union), Geneva, Switzerland.
- ITU-R. 2015. Method for the subjective assessment of intermediate quality levels of coding systems. ITU (International Telecommunication Union), Geneva, Switzerland.
- ITU-R. 2017. Methods for selecting and describing attributes and in the preparation of subjective tests. ITU (International Telecommunication Union), Geneva, Switzerland.
- Jekosch, U. 2005. Assigning meaning to sounds — semiotics in the context of product-sound design in *Communication Acoustics*, 193–221. ed. Jens Blauert. Berlin, Heidelberg, New York: Springer. https://doi.org/10.1007/3-540-27437-5_8.
- Joshi, D., R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Wang, J. Li, and J. Luo. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28 (5): 94–115. <https://doi.org/10.1109/msp.2011.941851>.
- Kahn, D. 2001. *Noise, Water, Meat: A History of Sound in the Arts*. Cambridge: The MIT Press.
- Kamaris, G., and J. Mourjopoulos. 2018. Stereo image localization maps for loudspeaker reproduction in rooms. In *Audio Engineering Society Convention 144*.
- Katz, R.A. 2015. *Mastering Audio: The Art and the Science*. Burlington: Focal Press.

- Kiebel, S.J., and K. Friston. 2001. Analysis of multisubject neuroimaging data using anatomically informed basis functions. *NeuroImage* 13 (6): 172.
- Konton, K.L., M. Megjhani, J.A. Brantley, J.G. Cruz-Garza, S. Nakagome, D. Robleto, M. White, E. Civillico, and J.L. Contreras-Vidal. 2015. Your brain on art: Emergent cortical dynamics during aesthetic experiences. *Frontiers in Human Neuroscience* 9.
- Kopco, N., V. Best, and B.G. Shinn-Cunningham. 2007. Sound localization with a preceding distractor. *The Journal of the Acoustical Society of America* 121 (1): 420–432. <https://doi.org/10.1121/1.2390677>.
- Leder, H., B. Belke, A. Oeberst, and D. Augustin. 2004. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology* 95 (4): 489–508. <https://doi.org/10.1348/0007126042369811>.
- Lepa, S., S. Weinzierl, H.-J. Maempel, and E. Ungeheuer. 2014. Emotional impact of different forms of spatialization in everyday mediated music listening: Placebo or technology effects? In *Audio Engineering Society Convention 136*.
- Levitin, D.J. 2011. *This Is Your Brain on Music: Understanding a Human Obsession*. London: Atlantic Books.
- Lindau, A., and S. Weinzierl. 2011. Assessing the plausibility of virtual acoustic environments. In *Forum Acusticum, European Acoustic Association, Aalborg, Denmark*, 1187–1192.
- Lund, T., and A. Mäkivirta. 2017. The bandwidth of human perception and its implications for pro audio. In *Audio Engineering Society Convention 143*.
- Martínez-Molina, N., E. Mas-Herrero, A. Rodríguez-Fornells, R.J. Zatorre, and J. Marco-Pallarés. 2016. Neural correlates of specific musical anhedonia. *Proceedings of the National Academy of Sciences* 113 (46): E7337–E7345.
- Mason, R. 2017. How important is accurate localization in reproduced sound? In *Audio Engineering Society Convention 142*.
- Mavridaki, E., and V. Mezaris. 2015. A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *IEEE International Conference on Image Processing (ICIP)*. <https://doi.org/10.1109/icip.2015.7350927>.
- McDermott, J.H. 2012. Auditory preferences and aesthetics: Music, voices, and everyday sounds. In *Neuroscience of Preference and Choice: Cognitive and Neural Mechanisms*, ed. R. Dolan, and T. Sharot. Amsterdam/New York: Elsevier/Academic.
- Merimaa, J., and V. Pulkki. 2005. Spatial impulse response rendering I: Analysis and synthesis. *Journal of the Audio Engineering Society* 53 (12): 1115–1127.
- Moon, P., and D.E. Spencer. 1944. Geometric formulation of classical color harmony. *Journal of the Optical Society of America* 34 (1): 46.
- Mourjopoulos, J. 2014. A paradigm shift for modeling sound sensation. In *40th International Computer Music Conference and 11th Sound and Music Computing Conference*.
- Munro, T., and R. Scruon. 2018. Aesthetics. <https://www.britannica.com/topic/aesthetics> (last accessed August 31, 2019).
- Nicol, R. 2020. Creating auditory illusions with spatial audio technologies. In *The Technology of Binaural Understanding*, ed. J. Blauert, and J. Braasch. Springer and ASA Press.
- Nishiyama, M., T. Okabe, I. Sato, and Y. Sato. 2011. Aesthetic quality classification of photographs based on color harmony. *IEEE Conference on Computer Vision and Pattern Recognition*. <https://doi.org/10.1109/cvpr.2011.5995539>.
- Pedersen, T., and N. Zacharov. 2015. The development of a sound wheel for reproduced sound.
- Posner, J., J.A. Russel, and B.S. Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology* 17 (3).
- Pulkki, V. 2001. Spatial sound generation and perception by amplitude panning techniques. Ph.D. thesis, Aalto University, Helsinki, Finland.
- Pulkki, V., S. Delikaris-Manias, and A. Politis. 2018. *Parametric Time-Frequency Domain Spatial Audio*. New York: Wiley.

- Raake, A., and J. Blauert. 2013. Comprehensive modeling of the formation process of sound-quality. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 76–81.
- Raake, A., and H. Wierstorf. 2014. A case for TWO!EARS in audio quality assessment. In *Forum Acusticum, Krakow, Poland*.
- Raake, A., and H. Wierstorf. 2020. Binaural evaluation of sound quality and quality-of-experience. In *The Technology, and of Binaural Understanding*, eds. J. Blauert and J. Braasch, 393–434. Cham, Switzerland: Springer and ASA Press.
- Rao, R.P.N., and D.H. Ballard. 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* 2 (1): 79–87.
- Reber, R., P. Winkielman, and N. Schwarz. 1998. Effects of perceptual fluency on affective judgments. *Psychological Science* 9 (1): 45–48. <https://doi.org/10.1111/1467-9280.00008>.
- Reber, R., N. Schwarz, and P. Winkielman. 2004. Processing fluency and aesthetic pleasure: Is beauty in the perceivers processing experience? *Personality and Social Psychology Review* 8 (4): 364–382. https://doi.org/10.1207/s15327957pspr0804_3.
- Redies, C. 2007. A universal model of aesthetic perception based on the sensory coding of natural stimuli. *Spatial Vision* 21 (1): 97–117.
- Redies, C. 2015. Combining universal beauty and cultural context in a unifying model of visual aesthetic experience. *Frontiers in Human Neuroscience* 09. <https://doi.org/10.3389/fnhum.2015.00218>.
- Renoult, J. 2016. The evolution of aesthetics: A review of models. In *Aesthetics and Neuroscience: Scientific and Artistic Perspectives*, ed. Z. Kapoula, and M. Vernet. Berlin: Springer.
- Renoult, J.P., J. Bovet, and M. Raymond. 2016. Beauty is in the efficient coding of the beholder. *Royal Society Open Science* 3 (3): 160027.
- Rumsey, F. 1998. Subjective assessment of the spatial attributes of reproduced sound. In *Audio Engineering Society Conference: 15th International Conference: Audio, Acoustics and Small Spaces*.
- Rumsey, F. 2002. Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society* 50 (9).
- Rumsey, F. 2008. Faithful to his master's voice? Questions of fidelity and infidelity in music recording. In *Recorded Music: Philosophical and Critical Reflections*, ed. M. Dogantan-Dack. London: Middlesex University Press.
- Rumsey, F. 2009. On the move with multichannel. *Journal of the Audio Engineering Society* 57 (10).
- Rumsey, F. 2011. Semantic audio: Machines get clever with music. *Journal of the Audio Engineering Society* 59 (11): 882–887.
- Rumsey, F. 2015a. Game audio: Generative music, emotions, and realism. *Journal of the Audio Engineering Society* 63 (4): 293–297.
- Rumsey, F. 2015b. Immersive audio, objects, and coding. *Journal of the Audio Engineering Society* 63 (5).
- Rumsey, F. 2016. Virtual reality: Mixing rendering, believability. *Journal of the Audio Engineering Society* 64 (12): 1073–1077.
- Rumsey, F. 2017. *Spatial Audio*. Boca Raton: CRC Press.
- Russel, J.A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39 (6).
- Schymura, C., and D. Kolossa. 2020. Blackboard systems for cognitive audition. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 91–111, Cham, Switzerland: Springer and ASA Press.
- Smith, S.R., and M.F. Bocko. 2017. Modeling the effects of rooms on frequency modulated tones. In *Audio Engineering Society Convention 143*.
- Sotujo, S., J. Thiemann, A. Kohlrausch, and S. Van de Paar. 2020. Auditory gestalt rules and their application. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 33–59, Cham, Switzerland: Springer and ASA Press.
- Stokes, D. 2009. Aesthetics and cognitive science. *Philosophy Compass* 4/5.

- Thiede, T., W.C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, and C. Colomes. 2000. PEAQ – the ITU standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society* 48 (1/2): 3–29.
- Tiihonen, M., E. Brattico, J. Maksimainen, J. Wikgren, and S. Saarikallio. 2017. Constituents of music and visual-art related pleasure – a critical integrative literature review. *Frontiers in Psychology* 8. <https://doi.org/10.3389/fpsyg.2017.01218>.
- Toole, F.E. 2018. *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms*. Florence: Routledge.
- Vickers, E. 2010. The loudness war: Background, speculation, and recommendations. In *Audio Engineering Society Convention 129*.
- Volk, C., S. Bech, T.H. Pedersen, and F. Christensen. 2015. Five aspects of maximizing objectivity from perceptual evaluations of loudspeakers: A literature study. In *Audio Engineering Society Convention 138*.
- Volk, C.P., S. Bech, T.H. Pedersen, and F. Christensen. 2017. Modeling perceptual characteristics of loudspeaker reproduction in a stereo setup. *Journal of the Audio Engineering Society* 65 (5): 356–366.
- Wierstorf, H., A. Raake, M. Geier, and S. Spors. 2013a. Perception of focused sources in wave field synthesis. *Journal of the Audio Engineering Society* 61 (1/2): 5–16.
- Wierstorf, H., A. Raake and S. Spors, 2013b. Binaural assessment of multichannel recordings. in: *The Technology of Binaural Listening*, ed. J. Blauert, Springer and ASA Press.
- Wikipedia. 2018. Aesthetics of music. http://en.wikipedia.org/wiki/Aesthetics_of_music (last accessed August 31, 2019).
- Williams, D. 2016. Toward emotionally-congruent dynamic soundtrack generation. *Journal of the Audio Engineering Society* 64 (9): 654–663. <https://doi.org/10.17743/jaes.2016.0038>.
- Zacharov, N., T. Pedersen, and C. Pike. 2016. A common lexicon for spatial sound quality assessment - latest developments. In *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*.
- Zacharov, N., C. Volk, and T. Stegenborg-Andersen. 2017. Comparison of hedonic and quality rating scales for perceptual evaluation of high- and intermediate quality stimuli. In *Audio Engineering Society Convention 143*.
- Zajonc, R.B. 1980. Feeling and thinking: Preferences need no inferences. *American Psychologist* 35 (2): 151–175. <https://doi.org/10.1037//0003-066x.35.2.151>.
- Zielinski, S., F. Rumsey, and S. Bech. 2008. On some biases encountered in modern audio quality listening tests-a review. *Journal of the Audio Engineering Society* 56 (6): 427–451.

A Virtual Testbed for Binaural Agents



Jens Blauert

Abstract Current developments in modeling the auditory system lead to increasing inclusion of cognitive functions, such as dynamic auditory scene analysis. This qualifies these systems as auditory front-ends for autonomous agents. Such agents can, for example, be mobile robotic systems, that is, they can move around in their environments, explore them, and develop internal models of them. Thereby, they can monitor their environments and become active in cases where potentially hazardous things happen. For example, in a Search-&-Rescue scenario (SAR), the agents could identify and save persons in dangerous situations. In this chapter, a virtual testbed for such systems is described that was developed in the EU project TWO!EARS (www.twoears.eu) There, in simulated scenarios, the agents have to localize and identify potential victims and, consequently, rescue them according to dynamic SAR plans. The actions are predominantly based on binaural cues, derived from the two ear signals of head-and-torso simulators (dummy heads) on carriages that can actively move about in the scenes to be explored. Such a simulation system can provide a tool to monitor and evaluate the cognitive processes of autonomous systems while these are dynamically executing assigned tasks.

1 Introduction

To qualify as acoustic front ends for autonomous agents, models of the auditory system need the capability of exploring and analyzing auditory scenes by using and interpreting acoustic cues. Auditory scene analysis (ASA) is the process by which the auditory system segregates the individual sounds in natural-world situations, whereby these sounds are usually spectrally and temporally interleaved and overlapping. Humans perform extremely well in such situations; that is, they can localize and comprehend multiple sound sources even in the presence of severe acoustic noise.

J. Blauert (✉)

Institute of Communication Acoustics, Ruhr-University, 44801 Bochum, Germany
e-mail: jens.blauert@rub.de

© Springer Nature Switzerland AG 2020

J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_17

For this kind of analysis, called auditory-stream segregation, the so-called *Gestalt*¹ rules are of relevance—compare Bregman (1990) and Jekosch (2005). As Sutojo et al. (2020), this volume, explain, human auditory systems use cognitive processes such as attention and prior knowledge when performing auditory scene analysis, rather than relying solely on the acoustic input to their two ears.

Transferring these human ASA skills to computer algorithms is the subject of *Computational Auditory Scene Analysis (CASA)*. For a comprehensive introduction to CASA systems see Wang and Brown (2006). Many CASA mechanisms use multi-microphone arrays—compare, for instance, Plinge et al. (2012) and EARS (2014). However, this technological approach differs significantly from the algorithms that are used in biological systems. Human beings did not develop multi-channel acoustic sensors but have to rely on just their two ears. Nevertheless, human performance in auditory scene analysis easily measures up to or even outperforms CASA systems in complex environmental settings.

The current chapter reports an example of an *intelligent CASA system* that incorporates cognitive processes as central elements. The system was developed in the context of the TWO!EARS project (www.twoears.eu [last accessed: September 1, 2019]). A mobile robotic agent, equipped with a head-and-torso simulator mounted on a mobile carriage, is controlled by a cognitive unit, the *blackboard system* (Schymura and Kolossa 2020), this volume, to explore acoustic scenarios actively. The resulting *active listening* or *dynamic auditory-scene analysis* comprises bottom-up data processing as well as top-down mechanisms connected by feedback loops. The architecture of this system is described in Blauert and Brown (2020), this volume.

The interleaving of bottom-up and top-down processes under the control of a blackboard system also constitute the basis for *attention-guidance* mechanisms, which are mandatory to cope with the vast amount of “information [that is] continuously available in the surrounding world” (Fabre-Thorpe 2003), and to achieve “real-time processing [performance] despite limited computational capacities” (Schauerte and Stiefelhagen 2013). Given a scene with significant auditory ambiguities, however, the above techniques might not be sufficient to untangle equivocal input information. Humans readily disambiguate such complex situations by resorting to visual information. To be able to measure up to this human skill, the robotic agent employed in TWO!EARS is realized as a multimodal-sensor platform that can resort to additional visual cues provided by an ego-centric camera system when necessary.

2 Audition in Cognitive Robotics

Robotic systems that rely on mobile agents for active exploration of the environment have recently caused significant research interest in robotics. While many of these systems are still restricted to visual information, some advanced approaches

¹The German term “Gestalt” describes an entity where the sum is perceived as more than the sum of its parts.

like the one of TWO!EARS incorporate auditory cues to enhance the robot's cognitive performance in complex scenarios. Compare, for example, the "iCub" robot (Metta et al. 2008; Ruesch et al. 2008) with *saliency maps* (Frintrop et al. 2010) for visual and acoustic input. Visually salient features include intensity, color hue, directional features, and motion. Auditory features are, among others, interaural level differences (ILDs) and interaural arrival-time differences (ITDs) to determine the azimuthal position of sound sources, and spectral notches to evaluate source elevation—compare Hörnstein et al. (2006). Saliency maps of both modalities are then combined by projecting them onto an *ego-sphere* (Ruesch et al. 2008), which is head-centered and fixed in relation to the robot's torso. In conjunction with a "dynamic-inhibition-of-return mechanism", the combined saliency maps allow the iCub robot to demonstrate a "rich attentional behavior" (Ruesch et al. 2008) and to autonomously explore multimodal stimuli in moderately complex environments.

Considering that purely visual exploration and attention-guidance systems fall short of reacting to salient events outside the visual field of view, Kuehn et al. (2012) learned from "Bayesian surprise techniques" (Itti and Baldi 2009) and introduced a concept of "auditory surprise" (Schauerte and Stiefelhagen 2013). Thereupon unexpected sound events are identified, and corresponding sound sources are localized using a "steered-response power [...] with phase transform [...] sound-source localization" approach—see Schauerte et al. (2011). Cue fusion then takes place based on a Gaussian-mixture model that integrates visual and auditory information in the sensor space. The proposed mechanism tries to generate exploration strategies to reduce the amount of necessary ego-motion for saving energy and, also, reduce "wear-and-tear" in the robotic device—see Kuehn et al. (2012).

Also, Okuno et al. (2001) emphasized the importance of audition in cognitive robotics. They created a multimodal control framework to guide a humanoid robot in service and assistance tasks (Kitano et al. 2000), based on a distributed architecture where vision, audition, motor control, and speech synthesis are realized as single modules that all communicate through a dedicated cognitive processing unit, the "association module" (Okuno et al. 2001). This module addresses tasks like associating audio streams with the corresponding visual streams and controls the robot's *focus-of-attention*. The system's audition component employs arrival-time and level differences of the signals captured by two microphones to perform sound-source localization in the horizontal plane (Nakadai et al. 2000).

In Walther and Cohen-L'hyver (2014), *dynamic weighting* methods after Cohen-L'hyver et al. (2015, 2020), this volume, are applied to perform computational auditory scene analysis in moderately complex acoustic scenarios. By continuously monitoring a given scene for incoming sounds, the system evaluates the "congruency" of a given stimulus. High weights are assigned to novel stimuli that seem to be incongruent with the current environmental model and thus appear potentially interesting. Lower weights are assigned to objects that have already been explored or are of lesser interest for the robot's actual task. In case of a novel sound signal with low congruency, the machine might instantaneously turn to this stimulus. Alternatively, it could deliberately suppress the turn-to-reflex in cases where the received acoustic input is congruent and thus less interesting.

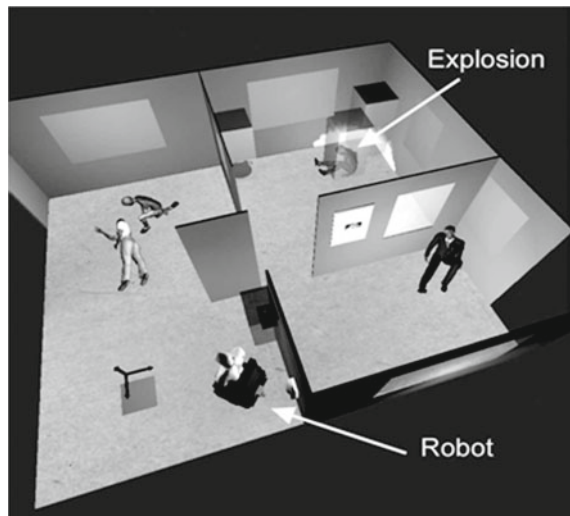
3 The Concept of the Virtual Testbed

As mentioned above, advanced models of the auditory system embody elements for cognitive processing and feedback loops that interlace bottom-up and top-down processes. For the developer and applicants of such systems, it is mandatory to understand what is going on inside the systems, in particular, when they perform complex tasks autonomously. To this end, specialized tools are necessary which enable continuous monitoring of the exploration, evaluation, decision, and task-planning processes in the systems. In the TWO!EARS project, a virtual testbed has been implemented for this purpose, with a focus on testing of intelligent models of the auditory system as prominent system components for autonomous robotic agents. The basic idea of the testbed design is to provide a virtual environment in which an (also virtual) mobile robotic agent can move about and perform tasks assigned to it.

A virtual testbed has the advantage that it can already be operational before the respective hardware system has been assembled. As an example, the provision of Search-&-Rescue (SAR) conditions for experimental tests is challenging in the real world, as such conditions may endanger human beings or the robot. A virtual testbed, however, can often avoid hardware expenses and significantly reduce testing-cycle time. Moreover, the virtual test environment can be designed to behave flexibly regarding new technologies, such as recent artificial-intelligence methods like *deep neural networks* and *reinforcement learning*—compare Goodfellow et al. (2016), Sutton (2018).

Following these lines of thinking, a demo scenario was modeled as a proof of concept—see Fig. 1. There, in a Search-&-Rescue (SAR) situation in a (moderately) complex environment, potential victims are localized, identified, and consequently rescued. The processes and actions are predominantly based on binaural cues, derived

Fig. 1 Search-&-Rescue (SAR) scenario used as an example for demonstrating the functionality of the virtual testbed of the TWO!EARS project. Rendered with Blender 3D-visualization software (Blender Foundation 2014)



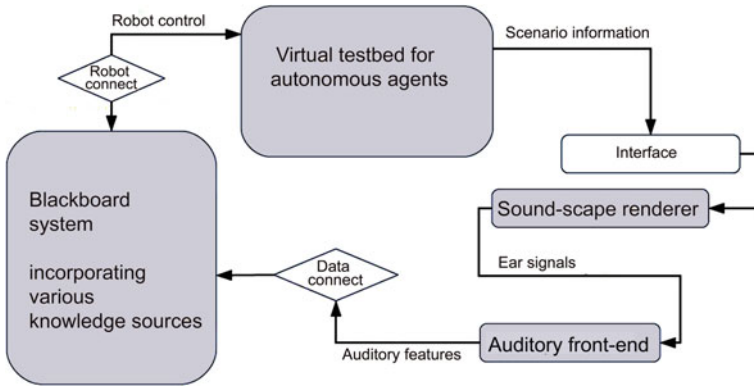


Fig. 2 Interplay of the virtual testbed and the blackboard system in a virtual audio-visual scenario. **Arrows** indicate the flow of information and symbolize issued control commands

from the ear signals of a virtual robotic platform that can actively move about in the scene to be explored. Visual cues are employed for assistance only if necessary. The action starts with a normal laboratory setting but suddenly evolves into a catastrophic situation, namely, after an explosion and fire breaking out in one corner of the laboratory. Thus, the attending persons turn into either victims or rescuers. The robotic agent enters the scenario and actively explores the terrain in order to infer the positions of all persons and then saves them successively in the order of their actual hazard score. The system newly assesses the individual hazard scores after each action step.

Technical details and a description of the algorithms of the TWO!EARS testbed are publicly available from the technical reports TWO!EARS (2015, 2016).

The virtual testbed is interleaved with a blackboard system which represents the core of the TWO!EARS system—compare, for example, Raake and Blauert (2013) or Blauert and Brown (2020), this volume. The advantages of blackboard structures, in general, are discussed in Schymura and Kolossa (2020), this volume. The architecture of the testbed/blackboard combination is depicted in Fig. 2 and explained below.

The left block of Fig. 2 symbolizes the blackboard. The blackboard incorporates a number of *knowledge sources* (KSs) and a *scheduler*. The knowledge sources are software modules with one specific functionality each. They define which data they need for execution and which data they produce. The blackboard system provides the tools for requesting and storing these data but does not care about their content. Each of the knowledge sources is in charge of a specific subtask that contributes to the solution of the general problem addressed. Knowledge sources are also called “experts”, as they act in a way similar to what human specialists would do. The scheduler is a software module that initiates knowledge sources to execute their respective subtasks. Thereby, it determines the order in which knowledge sources get executed, based on the current task and the data that are stored on the blackboard. The order of execution is rescheduled after every execution of a knowledge source.

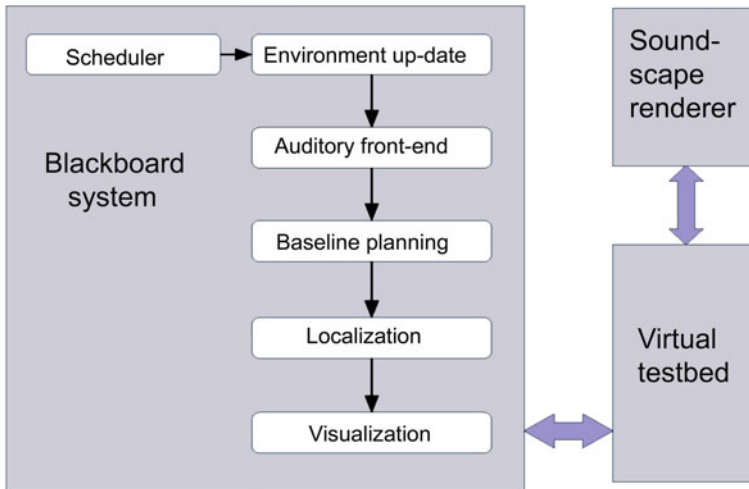


Fig. 3 Schematic of the blackboard architecture employed to perform multi-source localization based on a deep neural network (DNN). **Blueish arrows** indicate the flow of information and symbolize issued control commands. **Black arrows** illustrate the activation sequences of the depicted knowledge sources

The following sections provide short descriptions of knowledge sources and other software modules which are of particular relevance for the virtual testbed.

4 Exploration of the Scenario

The virtual scenario for testing autonomous agents is set up using a binaural mixing console. The mixing console has a data bank at its disposal with *head-related impulse responses* (HRIRs)² for various angles of sound incidence and sound-source distances such as the chosen test scenario requires. It generates the acoustic ear signals for the autonomous agents based on information of sound-source positions and emitted sound signals.³ There is a knowledge source that continually updates this information—the *UpdateEnvironmentKS*. The binaural mixing console interfaces with the next processing element, the *auditory front-end* via a special software module—the *SSRInterface*. The following explanations follow the schema depicted in Fig. 3.

The auditory front-end is the earliest processing stage of the TWO!EARS system. It provides bottom-up auditory signal processing as performed in the sub-cortical

²Head-related impulse responses (HRIRs) are the Fourier transforms of head-related transfer functions (HRTFs).

³In the current project the “SoundScape Renderer” (SSR) of Geier and Spors (2012) has been chosen for this purpose—see www.spatialaudio.net/ssr/ [last accessed: August 18, 2019].

stages of the human auditory system where the ear-input signals are transformed into multi-dimensional auditory representations. The output provided by this front-end consists of several transformed versions of ear signals enriched by perception-based descriptors, for example, interaural arrival-time differences (ITDs), interaural level differences (ILDs), interaural cross-correlation (IACC), signal onsets and offsets, loudness, pitch, rate maps, and binaural-activity maps.

An object-oriented approach is used throughout the system. This provides great flexibility and allows modification of bottom-up processing in response to feedback from higher levels of the system during run time. The auditory front-end supports online processing of the two-channel ear signals, and this is why it is used for the virtual testbed in the form of an *AuditoryFrontEndKS*.

4.1 Auditory-Object Localization

Perceptual objects are defined by their essential features, their position in space and time, their spatial extent, and their relation to other objects. In other words, they exist at a certain time at a certain locus. Sound-source localization or, to be more specific, the determination of the positions of auditory objects in the perceptual space is a basic requirement for any auditory-object formation—refer to Blauert (1997) for fundamentals of auditory localization.

In the virtual testbed reported here, sound-source localization is accomplished by a *deep neural network* (DNN). Networks of this kind have proven successful for the determination of the directions of multiple sound sources, such as concurrent speakers—even in noisy and reverberant environments. The system of Ma et al. (2015) is an example, in which a DNN is used to learn the relationship between the source azimuth and binaural cues, namely, interaural cross-correlation and interaural level differences of the signals arriving at the two ears. The DNN was trained using a multi-condition approach, that is, spatially diffuse noise was added to the training signals at different signal-to-noise ratios to improve robustness to reverberation. The authors show that their system can accurately localize target sources in challenging conditions—even when concurrent sound sources and room reverberation are present.

However, there is one additional complication that had to be addressed, that is, the following. Auditory localization was traditionally discussed as a static phenomenon, not considering that the ears are positioned on the head which is movable in six degrees of freedom. The exploitation of additional cues as collected by head movements improves the localization capabilities considerably, for instance, in the course of exploring unknown environments—compare Braasch et al. (2013) and Blauert and Brown (2020), Pastore et al. (2020), both this volume. In particular, head movement is needed to solve directional ambiguities such as front-back confusion and to move the head into a suitable position for the segregation of desired signal components from undesired ones, such as noise, reverberation, and/or concurrent talkers (Braasch et al. 2011, 2013). For this reason, the virtual agent in the current system is enabled to execute slow rotations about its vertical axis when necessary—angular velocity 15°/s.

To evaluate the auditory localization capabilities of the virtual testbed, the following scenario was generated. From the “real” facility, the ADREAM apartment, which was used for tests with the “real” autonomous agent, a virtual copy was set up (ADREAM 2014). Up to seven virtual sound sources were placed at different positions inside the apartment, namely, five human beings, a dog, and an open fire. The sound sources emit signals that include human voices yelling for help, a siren, a barking dog, and the fire. Each source displays intermitting activation with artificial silence intervals of 0.5 s in addition to the natural silence intervals contained in each of the sound signals anyhow. In this scenario, the virtual robot had to infer the azimuths of all sound sources.

As ground-truth information, the robot was supplied with a floor plan of the apartment with all sound-insulating walls charted.⁴ The motion behavior of the robotic agent while localizing the sound sources is controlled by a dedicated planning module. In the course of the localization experiment, the following behavior was observed.

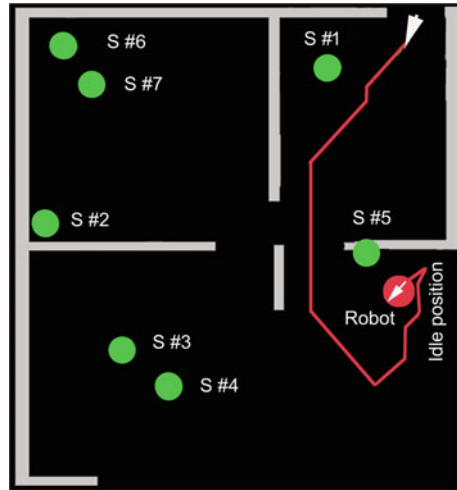
To localize the auditory objects, the virtual agent roams the virtual scenario along a path proposed by the planning module. While following this path, the robot performs a localization attempt by listening for any emitted sounds and infers the azimuths of all discriminable active sound sources by feeding data from the auditory front end into the *DnnLocationKS*. This knowledge source, in turn, generates a discrete distribution, that yields the probability of sound-source presence over head-centric azimuth in each simulation frame.

Using the current head orientation of the agent, as provided by the virtual testbed, the head-centric coordinates are transformed into world coordinates. In the transformed probability distribution, $p_1(\varphi)$, peaks are popping up in the most likely positions of auditory objects. However, in a plot of this distribution, (mirror) peaks may develop at positions roughly symmetric to the ear axis. This is the result of front-back confusions. This problem is solved by head rotation. From a head position shifted in azimuth concerning the initial head position, another probability distribution, $p_2(\varphi)$, is generated. By adding up the two distributions, $p_1(\varphi)$ and $p_2(\varphi)$, one of each pair of symmetric peaks, the “ghosts”, level out and eventual front-back-confusions are thus disambiguated.

Figure 4 depicts the floor plan of the test scenario with all seven estimated source positions marked. Further plotted are a roaming path and the final position of the autonomous agent. From the latter position, it monitors the scenario for any new things to happen in *idle mode*.

⁴Sound-reflections from the walls were not considered, as this did not appear to be of importance for the current localization task. If this became necessary for tests in more complex scenarios, a precedence-effect processor had to be implemented—such as the one described by Braasch (2020), this volume.

Fig. 4 Floor plan of the test scenario with the detected positions of sound-sources, a roaming path and the idle position of the agent



4.2 Auditory-Object Identification

As mentioned above, perceptual objects are defined by more than their position in space and time. Further essential features are, for example, psychoacoustics- and sensory-psychology-related attributes and physical attributes. Further, most importantly, the *meaning* that the objects convey and the specific functionality that they stand for. Autonomous agents must develop an idea of these object characteristics, at least to the extent which is relevant for the tasks that are assigned to them. An important process to this end is what is known as the “*binding*”—compare von der Malsburg (1999). This is the mutual allocation of those features which constitute the “*identity*” of a perceptual entity on the one hand and its location in time and space on the other hand. In the virtual testbed described here, the binding process is accomplished by a specific knowledge source—the *BindingKS*.

An important reason for building the virtual testbed at all was that it allowed the testing of procedures that were not yet available with the physically “real” autonomous agent in the course of the development of the TWO!EARS project. For instance, some data needed for performing the binding were still lacking at this time. However, in the virtual testbed, they could be emulated by integrating ground-truth information from the virtual scenario. As a positive side effect, experiments run significantly faster in emulated scenarios, because auralization and auditory-feature extraction are bypassed. Further, the emulation supports setting up complex experiments easily and fast that could not readily be realized otherwise.⁵ The following explanations follow the schema depicted in Fig. 5

⁵If at a later point in time a sufficient amount of experimental data from the “real” agent is available, the emulated identity labels can be replaced by real ones. Identity classes could even be compiled automatically from experimental data.

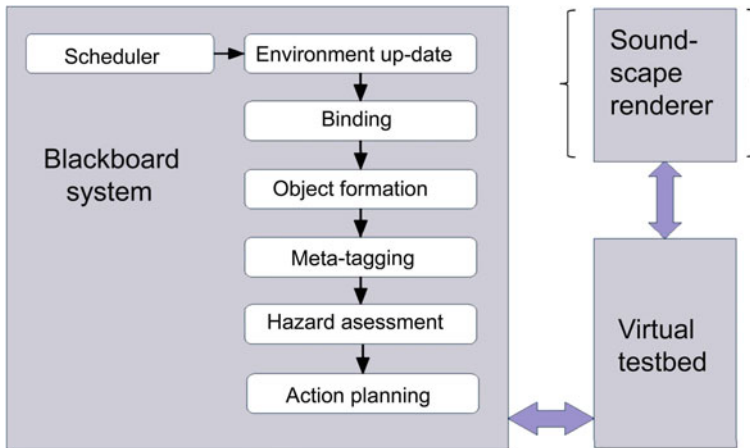


Fig. 5 Processing sequence in the virtual testbed for auditory object identification and subsequent action planning. **Blueish arrows** indicate the flow of information and symbolize issued control commands. **Black arrows** mark the activation sequence of the knowledge sources. Note that the sound-scape renderer, in **curled brackets**, is not active in the *emulation mode* of the testbed

Binding

As stated above, the “Wheres” and “Whats” of all perceived sound sources have to be combined by appropriate binding mechanisms for successful auditory-object formation. To that end, the *BindingKS* imports ground-truth-scenario information from the core of the virtual testbed and generates, on a per-frame basis, a set of *binding hypotheses* corresponding to all currently active sound sources. Given the location estimates for all active sources in each frame, it becomes straightforward to augment those estimates with the corresponding ground-truth labels. The *BindingKS* then forwards the resulting structure to the blackboard, thus making it available to downstream knowledge sources.

The basic concept of the *BindingKS* is explained in the following, for details of the algorithm see TWO!EARS (2016, pp. 92–93). The *BindingKS* builds on the following data.

- Azimuthal-plane-based ground-truth positions of all sound sources in the given scenario
- Ground-truth *labels* which represent the respective *identities* of all auditory percepts in the scenario
- Head poses of the autonomous agent in any given frame—including head orientation.

Object Formation

Binding is subsequently followed by the *formation of auditory objects*. The Auditory-Object-Formation Knowledge Source, *AuditoryObjectFormationKS*, is in charge of

this step. Based on the source-position estimates of the *DnnLocationKS*, the robotic agent infers and refines the *x/y*-plane position of each sound source in each frame. The acuity of the position estimates gradually increases with increasing emulation time, as information from more frames aggregates and cancels out uncertainties in azimuth estimation for the robot’s self-localization mechanism. To that end, the machine switches to *patrol mode* and roams the given scenario. In doing so, it uses a simplistic path-planning scheme that integrates basic collision avoidance—modified from Premakumar (2016). Patrolling is continued until all source positions are determined with sufficient precision. During a patrol, the robotic agent maintains a set of *auditory-object hypotheses*. Each of them represents a self-contained, expandable set of data that models the robot’s knowledge of each acoustically observed sound source in each frame.

The information contained in auditory-object hypotheses grows and adapts while the robot moves along the prescribed patrolling path. Simultaneously the localization acuity increases, and the variance, v , of the framewise collected positions in the horizontal plane decreases. The averages of the variances—over ten simulation frames and across all object hypotheses—is called the “*average global-position uncertainty*”, \tilde{v} . Once a preset threshold for this quantity, say $\tilde{v} \leq 0.01$, is undercut, patrolling is stopped, and the next processing step is initiated.

Eventually, the *AuditoryObjectFormationKS* augments each auditory object hypothesis with its individual position variance. In addition, this knowledge source stores \tilde{v} in the blackboard for further processing by downstream system blocks.

Meta-tagging

As mentioned above, the virtual testbed focuses on the cognitive domain and operates on the symbolic level. *Meta information* required herein is provided by the *AuditoryMetaTaggingKS*. This knowledge source augments each auditory object hypothesis with additional *meta tags* that define the abstract characteristics of the corresponding sound sources in a given emulation frame. Note that the metadata used here are purely emulated. The advantage of such a procedure is that it allows for performing cognitive experiments of increased complexity, even if the required meta information is currently not yet available from lower stages of the TWO!EARS framework. The available meta tags as listed in Table 1 are generated by the *AuditoryMetaTaggingKS*.

Table 1 Meta-tags assigned to object hypotheses in the virtual testbed

Meta class	Meta subclass
Category	Human, Animal, Threat, Alert
Role	Employee, Rescuer, Victim, Fire, Siren, Dog
Gender	Male, female, n.a.
Stress	Categorized individually
Loudness	Categorized individually
Age	Categorized individually

In this table, “n.a.” indicates that the corresponding meta tag may not be applicable in the specific case. For instance, it would be pointless to assign stress values to an auditory object of role “fire”, or supply an auditory object of role “siren” with “age” information. Be reminded that the virtual testbed contains ground-truth meta-information for all instantiated sound sources. Thus, each auditory-object hypothesis can principally be supplied with perfect meta-knowledge of the emulated environment. This would never be possible in real-world scenarios, as meta information can only be extracted via noisy sensors, consequently resulting in an imperfect assignment of data to the corresponding auditory percepts.

To account for such sensor noise in the existing virtual testbed, *membership scores* for all classes defined in Table 1 are calculated by artificially degrading the ground-truth information by adding noise to the acoustic signals.

Hazard Assessment

Given the above meta information, the virtual testbed calculates individual *hazard scores*, H , for all the auditory objects. This is done in the *HazardAssessmentKS* knowledge source. In a first step, a *rescue score*, R , is appointed to each object, referring to the most probable meta-classes that the respective object belongs to in terms of category, role, and gender—determined for each simulation frame.

High values of the rescue score may induce ambiguous cases. For instance, an object that relates to a particular sound source is a rescuer itself. Thus it can be assumed that only minor help from the robotic agent is needed, if at all. In a second case, an observed object is positioned closely to another object that has taken on the role of a “rescuer”. Then the system expects the nearby rescuer to look after the observed object. The robot will instead focus its attentional resources on entities with lower rescue scores. Note that in both cases, an increase of the rescue score causes the hazard score to decrease.

The rescue score is countered by a *threat score* for the individual objects. This takes on high values if the observed entity is expected to be positioned close to another one that likely belongs to the “threat” category. In such a dangerous situation, the attention of the robot has to focus on the threatened entity and thus increases its hazard score. The hazard score will also increase if the observed entities belong to the following classes.

- Is likely to be a victim
- Shows increased voice stress or loudness
- Is close to a threat, for example, a fire.

During rescue attempts, all animate beings have to be evacuated from the scenario. However, as humans have to be rescued first, their individual threat score is post-processed to confirm their priority to be rescued. It is further assumed that inanimate entities cannot be threatened. Thus, their threat scores are set to zero.

The above-mentioned heuristic, relatively simple assessment rules for the individual hazard scores resulted in a reasonable behavior of the robotic agent in the SAR experiment, which has been set up as a proof of the basic concept of the virtual testbed.

The individual hazard scores, H , were derived from weighted algebraic sums of the average global-position uncertainty, \bar{v} , the rescue score, R , and a loudness index, L . The individual rescue and hazard scores were averaged over a sliding time window of the last 30 frames where applicable, resulting in *smoothed individual-hazard-scores*. All individual rescue and hazard scores, as well as the global ones, are communicated to the blackboard system for use in further knowledge sources, for example, in the *PlanningKS*. The derivation of hazard scores can readily be extended to suit the demands of more complex scenarios. For upcoming system versions, it would also be possible to have human assessors judge on hazard scores in a variety of emulated dynamic ASA situations. Data recorded from these trials would then be used to train, for example, a neural network that infers the hazard score directly and replaces the ad-hoc solution proposed above.

5 Action Planning

The knowledge sources defined above act together in order to allow for *bottom-up* scenario analysis. However, without *top-down* active exploration, the *AuditoryObjectFormationKS* would not be able to localize overheard sound sources. In turn, computation of the hazard scores as provided by the *HazardAssessmentKS* would be flawed. Consequently, it would be impossible to derive a meaningful action plan for the robotic rescuer without adequate scenario understanding driven by top-down mechanisms.

To account for these insights, the blackboard architecture is augmented with a cognitive expert subsystem, the *PlanningKS*. This knowledge source enables meaning assignment and scene understanding together with high-level planning as well as active scenario exploration. The *PlanningKS* can, in a way, be seen as the “brain” of the autonomous robotic agent.

Realized as a task stack, the *PlanningKS* employs a manually derived rule set to issue new tasks and provide meaningful robotic behavior in moderately complex scenarios.

To be sure, the internal architecture of the *PlanningKS* has to be adaptable to novel situations. For instance, the understanding and handling of the demo SAR scenario introduced in Sect. 3 is enabled through the rules and tasks encoded in Fig. 6. Only the major tasks have been depicted in this figure; minor subtasks like actuator control have been left out for clarity.

In future system versions, the manual adaptation of the *PlanningKS* could be automatized as follows. On the one hand, neural-network methods could, for instance, be used to infer purposeful robotic action plans directly from data collected in human trials. On the other hand, application of reinforcement-learning techniques could be employed to enable the robot to discover reasonable action patterns in a completely autonomous manner.

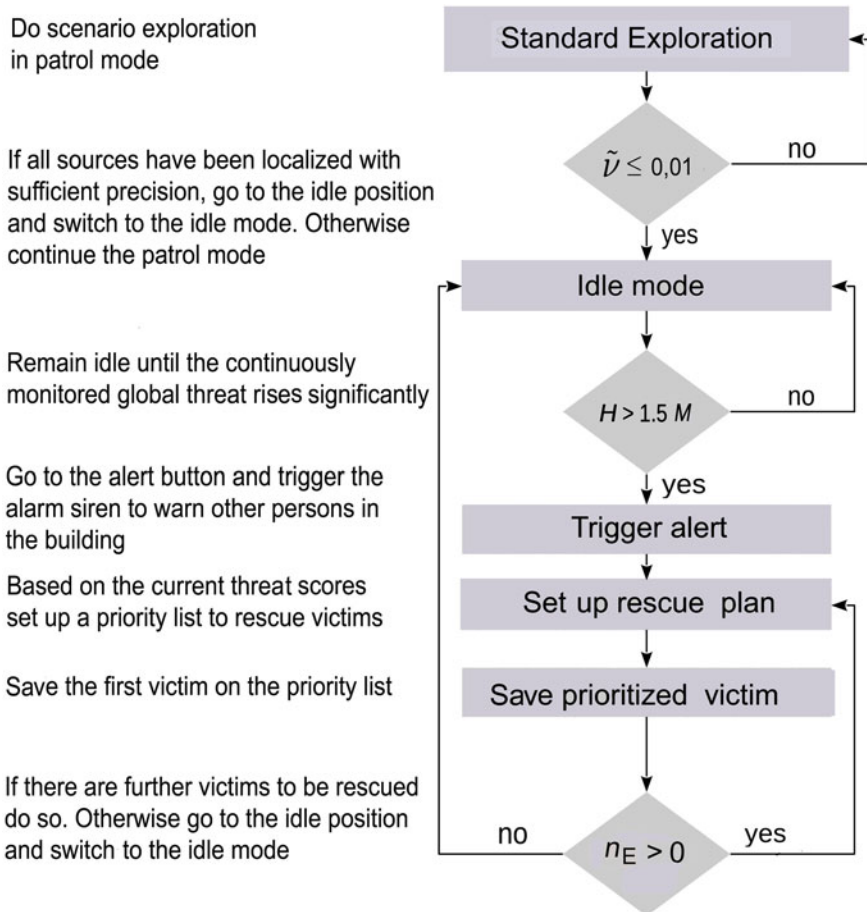


Fig. 6 Overview of the *PlanningKS* architecture described in Sect. 5. The flow plan on the right subsumes the tasks (rectangles) and decision rules (diamonds) embodied in the framework for scene understanding and active exploration. The hints on the left provide details of the corresponding diagram blocks. \tilde{v} ... average global-position uncertainty, H ... individual hazard score, M ... average global hazard score, n_E ... number of remaining victims

6 Multimodal Cue Integration

As stated above, the TWO!EARS framework is aimed at multimodal augmentation of auditory scene understanding. To this end, the physical robotic agent is equipped with a binocular camera system that enables the capturing of video footage in given scenarios. Visual cues extracted from the image streams that the cameras deliver can be used to complement incoming auditory information, thereby enhancing the robot’s comprehension of the explored environment. The virtual robotic agent is

Table 2 Meta characteristics of the entities present in the cognitive experiment discussed in Sect. 7

Entity	Category	Pre-event role	Post-event role	Gender	Age
S #1	Human	Employee	Victim	Male	25
S #2	Animal	Dog	Victim	Male	2
S #3	Human	Employee	Rescuer	Female	30
S #4	Human	Employee	Victim	Male	40
S #5	Alert	Siren	Siren	n.a.	n.a.
S #6	Threat	Fire	Fire	n.a.	n.a.
S #7	Human	Employee	Victim	Female	20

also equipped with a monocular camera to assess the benefits of audio-visual cue integration within the emulation framework. This camera allows capturing the robot’s field of view.

The entities enrolled in the given scenario correspond to a subset, namely, sources S #1, S #3, S #4, and S #7—compare Fig. 4—embracing the sources defined in Table 2, with their individual roles switched to “victim”. All entities are in a panic, causing nearly identical stress levels. The procumbent female victim, S #7, is supposed to be severely injured, with her utterances significantly muffled. In contrast, the entities related to S #1 and S #3 appear to be physically integer and are assumed to yell for help actively. Thus one gets similar stress levels for all emulation frames.

Since there are no rescuers or threats present in the given scene, the initial hazard scores computed for the entities are defined as directly proportional to their individual loudness levels. As a consequence, the robot would at first evacuate the active, intact entities corresponding to sound sources S #1 and S #3, disregarding the helpless persons represented by S #4 and S #7. Such behavior, however, would be in contrast to human intuition and thus clearly inadmissible.

Incoming visual information is exploited in the following way to suppress such inadequate behavior. A *histogram of oriented gradients* (Dalal and Triggs 2005) from the *OpenCV* library of WillowGarage (2014) is used to detect persons in upright posture. With this additional information, the robot’s rescuing pattern as defined by the *PlanningKS* is adapted. Once the robot has determined the positions of all sound sources with sufficient reliability—solely based on acoustic cues—it activates its camera and focuses sequentially on the estimated individual source locations. The assumed *physical integrity* of the victim corresponding to sound source S #3 can thus be checked based on visual cues—see Fig. 7.

As a result, if the robot’s heading is geared towards S #3, and the position detector reports the visual presence of an upright person. It is concluded that the focused victim is fully conscious and physically integer. On the contrary, if the detector displayed a negative response, the victim at S #3 would be deemed severely injured and probably dizzy. The computational load induced by the position detector is kept at bay by shutting down the camera of the robot unless visual augmentation is actually demanded.

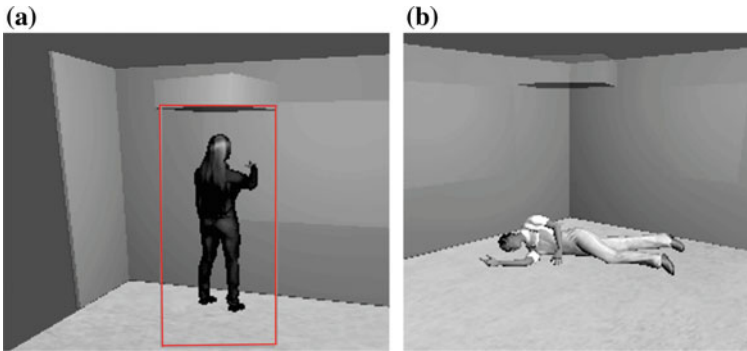


Fig. 7 Visual assessment of the physical integrity of emulated victims with a position detector. **a** A victim has been focused, the position detector reports the presence of an upright person (**red box**). Consequently it is inferred that this person is fully conscious and physically integer. **b** Another victim has been focused, yet the detector shows a negative response. Thus, this person is likely to be injured and probably unconscious or dizzy

The virtual testbed memorizes the physical-integrity values for all assessed entities and uses them to update the characteristics of the corresponding sound sources. In case of entities classified as injured, their hazard score is significantly increased, causing helpless entities to become prioritized in the actual rescue plan. Accordingly, the procumbent person corresponding to S#4 will be evacuated with priority, resulting in the behavior of the robotic agent that matches human intuition.

7 Test-Run of the Virtual Testbed

As a demonstration of the basic functionality of the auditory virtual testbed in a dynamic auditory scene, a test-run in the simplified scenario as depicted in Fig. 4 was performed. The emulation of the scenario in the virtual testbed has a duration of 400 s. It starts in normal lab conditions. Then, after $T=60$ s, the situation evolves into a catastrophic scenario. After an assumed explosion, the attending laboratory employees turn into either victims or rescuers, and a fire starts in one corner of the laboratory. Table 2 subsumes the meta-characteristics of all entities present in the proposed scenario, including their roles before and after the explosion.

Each of the entities in the table above corresponds to an emulated sound source, distinguished by an individual *utterance schedule* that contains the emission pattern and potential role changes for each entity. In the *emulation mode*, the virtual testbed does not require the availability of physical stimuli connected to the utterances stored. This allows defining entities of nearly arbitrary category and role without the need for huge sound databases. Note further that emission and silence intervals in the utterances are superposed by stochastic noise with preset levels.

Assume that before the explosion all animate entities display a low stress level, indicated by their vocal activities. The emulated sound sources corresponding to the “fire” and “siren” entities remain inactive. After the explosion, the stress level suddenly rises in all animate entities. S#1 is assumed to become unconscious, ceasing its acoustic emission after a few seconds. S#4 is expected to be severely injured, causing the loudness level of its utterances to decrease significantly. Note that a fire-alarm siren is available, but its activation is deliberately postponed until the autonomous agent triggers it by pressing the “trigger alert” button. The robotic agent acts with respect to the schedule defined in Fig. 6 to sense upcoming catastrophic conditions and, eventually, evacuate all animate entities from the scene.

The virtual testbed allows to automatically generate a range of different scenarios with varying characteristics, thus allowing for quantitative assessment of the performance of SAR schemes encoded in the *PlanningKS*. Focusing on the SAR strategy discussed in Fig. 6, modified scenarios were generated by randomly altering the x/y -positions of all animate entities. “Forbidden areas” were defined where no animate entity was placed. In this way, the randomly positioned sources are kept away from walls and become unlikely to stall the robot during scenario exploration completely.

In the scenario shown in Fig. 1, the time span for the emulation plus that required for the evacuation of all animate entities from the scene was 300 s on average, with a standard deviation of 38 s. In upcoming experiments, these values will have to be compared to results from trials where human assessors guide the robotic agent manually through numerous emulated rescue attempts. This would also set the pace for perceptual evaluation in addition to the instrumental one applied so far.

8 Discussion and Conclusion

It is common practice in science and technology to provide virtual environments for actions which are too expensive, too complex, too slow to realize, or too dangerous to be performed in corresponding “real” environments. In this chapter a *virtual testbed* is described in which a mobile *virtual autonomous robotic agent* acts in a *virtual scenario*. The unique feature of the testbed is that the agent autonomously explores its environment predominantly based on auditory cues, as are derived from the input signals to the two ears of a head-and-torso simulator mounted on a mobile platform—all virtual! Complementary visual cues are only used in rare cases where auditory information is not sufficient for the task assigned to the agent. The virtual testbed has been developed in the course of an international research project to study the scientific problems that the realization of such a system would involve. When the virtual test reported here was developed, the situation regarding relevant experimental data was scarce. Therefore, in many knowledge sources, heuristic-rule sets had to be employed. Once available data sets are capacious enough for automatic analysis, the heuristic rules may be replaced by statistics-based end-to-end classifiers. In any case, the basic feasibility of virtual auditory testbeds could be demonstrated.

Acknowledgements The work reported in this chapter has received funding from the European Union 7th Framework Program (FP7/2007-2013, grant agreement No.618075, project acronym TWO!EARS). The results have been achieved in a working group composed of B. Cohen-L'hyver, Ch. R. Kim, H. Wierstorf, Y. Kashef, J. Mohr, J. Braasch, N. Ma, S. Argentieri, G. Bustamante, P. Danès, Th. Walther, Th. Fogue, A. Podlubne, T. May, Y. Guo, and the current, reporting author. Part of the material dealt with refers to public technical reports delivered to the European Commission TWO!EARS (2015, 2016). The software is publicly available from the project's website. Thanks go to three anonymous reviewers for valuable comments and suggestions.

References

- ADREAM. 2014. Lab. for analysis and architecture of systems, F–Toulouse. <https://www.laas.fr/public/en/adream>. Last accessed 18 Aug 2019.
- Blauert, J. 1997. *Spatial Hearing—The Psychophysics of Human Sound Localization*, 2nd ed. Cambridge, MA: The MIT-Press (expanded and revised edition of Räumliches Hören, S. Hirzel, Stuttgart, 1974).
- Blauert, J., and G. Brown. 2020. Reflexive and reflective auditory feedback. In *The Technology, and of Binaural Understanding*, eds. J. Blauert and J. Braasch, 3–31. Cham, Switzerland: Springer and ASA Press.
- Blender Foundation. 2014. Blender-3D open source animation suite. <http://www.blender.org/>. Last accessed 18 Aug 2019.
- Braasch, J. 2020. Binaural modeling from an evolving-habitat perspective. In *The Technology, and of Binaural Understanding*, eds. J. Blauert and J. Braasch, 251–286. Cham, Switzerland: Springer and ASA Press.
- Braasch, J., S. Clapp, A. Parks, T. Pastore, and N. Xiang. 2013. A binaural model that analyses aural spaces and stereophonic reproduction systems by utilizing head movements. In *The Technology of Binaural Listening*, ed. J. Blauert, 201–223. Springer and ASA Press.
- Braasch, J., A. Parks, and N. Xiang. 2011. Utilizing head movements in the binaural assessment of room acoustics and analysis of complex sound source scenarios. *The Journal of the Acoustical Society of America* 129: 2486.
- Bregman, A. 1990. *Auditory Scene Analysis—The Perceptual Organization of Sound*. Cambridge, MA: The MIT Press.
- Cohen-L'hyver, B., S. Argentieri, and B. Gas. 2015. Modulating the auditory Turn-to-Reflex on the basis of multimodal feedback loops: The Dynamic Weighting Model. In *IEEE Robio 2016—International Conference on Robotics and Biomimetics*.
- Cohen-L'hyver, B., S. Argentieri, and B. Gas. 2020. Audition as a trigger of head movements. In *The Technology, and of Binaural Understanding*, eds. J. Blauert and J. Braasch, 697–731. Cham, Switzerland: Springer and ASA Press.
- Dalal, N., and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 886–893.
- EARS 2014. Embodied audition for robots. <https://robot-ears.eu/>. Last accessed 18 Aug 2019.
- Fabre-Thorpe, M. 2003. Visual categorization: Accessing abstraction in non-human primates. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* 358: 1215–1223.
- Frintrop, S., E. Rome, and H.I. Christensen. 2010. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception* 7 (1): 6:1–6:39.
- Geier, M., and S. Spors. 2012. Spatial audio reproduction with the soundscape renderer. In *27th Tonmeisterstagung—VDT International Convention*.
- Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. Cambridge, MA; GB, London: The MIT Press.

- Hörnstein, J., M. Lopes, J. Santos-Victor, and F. Lacerda. 2006. Sound localization for humanoid robots—Building audio-motor maps based on the HRTF. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1170–1176.
- Itti, L., and P. Baldi. 2009. Bayesian surprise attracts human attention. *Vision Research* 49 (10): 1295–1306.
- Jekosch, U. 2005. Assigning of meaning to sounds—Semiotics in the context of product-sound design. In *Communication Acoustics*, ed. J. Blauert, 193–221. Springer.
- Kitano, H., H.G. Okuno, K. Nakadai, T. Sabisch, and T. Matsui. 2000. Design and architecture of SIG the humanoid: An experimental platform for integrated perception in RoboCup humanoid challenge. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 181–190.
- Kuehn, B., B. Schauerte, K. Kroschel, and R. Stiefelhagen. 2012. Multimodal saliency-based attention: A lazy robot’s approach. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 807–814.
- Ma, N., G.J. Brown, and T. May. 2015. Robust localisation of multiple speakers exploiting deep neural networks and head movements. In *Proceedings of Interspeech15*, 2679–2683.
- Metta, G., G. Sandini, D. Vernon, L. Natale, and F. Nori. 2008. The iCub humanoid robot: An open platform for research in embodied cognition. In *Proceedings of 8th Workshop Performance Metrics for Intelligent Systems*, 50–56.
- Nakadai, K., T. Lourens, H.G. Okuno, and H. Kitano. 2000. Active audition for humanoid. In *Proceedings 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence*, 832–839.
- Okuno, H.G., K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano. 2001. Human-robot interaction through real-time auditory and visual multiple-talker tracking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1402–1409.
- Pastore, T., Y. Zhou, and A. Yost. 2020. Cross-modal and cognitive processes in sound localization. In *The Technology, and of Binaural Understanding*, eds. J. Blauert and J. Braasch, 315–350. Cham, Switzerland: Springer and ASA Press.
- Plinge, A., M.H. Hennecke, and G.A. Fink. 2012. Reverberation-robust online multi-speaker tracking by using a microphone array and CASA processing. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*.
- Premakumar, P. 2016. A* (A star) search path planning tutorial. <https://de.mathworks.com/matlabcentral/fileexchange/26248-a-a-star-search-for-path-planning-tutorial>. Last accessed 18 Aug 2019.
- Raake, A., and J. Blauert. 2013. Comprehensive modeling of the formation process of sound-quality. In *5th International Workshop Quality of Multimedia Experience (QoMEX, Klagenfurt)*, 76–81.
- Ruesch, J., M. Lopes, A. Bernardino, J. Hörnstein, J. Santos-Victor, and R. Pfeifer. 2008. Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *IEEE International Conference on Robotics and Automation*, 962–967.
- Schauerte, B., B. Kühn, K. Kroschel, and R. Stiefelhagen. 2011. Multimodal saliency-based attention for object-based scene analysis. In *IEEE International Conference on Intelligent Robots and Systems*, 1173–1179.
- Schauerte, B., and R. Stiefelhagen. 2013. “Wow!” Bayesian surprise for salient acoustic event detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6402–6406.
- Schymura, C., and D. Kolossa. 2020. Blackboard systems for modeling binaural understanding. In *The Technology, and of Binaural Understanding*, eds. J. Blauert and J. Braasch, 91–111. Cham, Switzerland: Springer and ASA Press.
- Sutojo, S., S. Van de Par, J. Thiemann, and A. Kohlrausch. 2020. Auditory Gestalt rules and their application. In *The Technology, and of Binaural Understanding*, eds. J. Blauert and J. Braasch, 33–59. Cham, Switzerland: Springer and ASA Press.
- Sutton, R. 2018. *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA: The MIT Press.

- TWO!EARS. 2015. Specification of feedback loops and implementation progress. In *Two!Ears Publications*, ed. J. Blauert and T. Walther, Chap. Project deliverables, item d4.2, pp. 56–61, <https://doi.org/10.5281/zenodo.2595224>.
- TWO!EARS 2016. Final integration-&-evaluation. In *Two!Ears Publications*, ed. J. Blauert and T. Walther, Chap. Project deliverables, item d4.3. <https://doi.org/10.5281/zenodo.2591202>.
- von der Malsburg, C. 1999. The what and why of binding: The modeler's perspective. *Neuron* 24: 95–104.
- Walther, T., and B. Cohen-L'hyver. 2014. Multimodal feedback in auditory-based active scene exploration. In *Proceedings of Forum Acusticum*. Kraków, Poland.
- Wang, D., and G.E. Brown. 2006. Computational auditory scene analysis: Principles, algorithms, and applications. *IEEE Xplore*. <https://ieeexplore.ieee.org/document/4429320>. Last accessed 18 Aug 2019.
- WillowGarage 2014. Open source computer vision library. <https://ieeexplore.ieee.org/document/4429320>. Last accessed 18 August 2019.

Binaural Technology for Machine Speech Recognition and Understanding



Richard M. Stern and Anjali Menon

Abstract It is well known that binaural processing is very useful for separating incoming sound sources as well as for improving speech intelligibility in reverberant environments. This chapter describes and compares a number of ways in which automatic-speech-recognition accuracy in difficult acoustical environments can be improved through the use of signal processing techniques that are motivated by our understanding of binaural perception and binaural technology. These approaches are all based on the exploitation of interaural differences in arrival time and intensity of the signals arriving at the two ears to separate signals according to direction of arrival and to enhance the desired target signal. Their structure is motivated by classic models of binaural hearing as well as the precedence effect. We describe the structure and operation of a number of methods that use two or more microphones to improve the accuracy of automatic-speech-recognition systems operating in cluttered, noisy, and reverberant environments. The individual implementations differ in the methods by which binaural principles are imposed on speech processing, and in the precise mechanism used to extract interaural time and intensity differences. Algorithms that exploit binaural information can provide substantially improved speech-recognition accuracy in noisy, cluttered, and reverberant environments compared to baseline delay-and-sum beamforming. The type of signal manipulation that is most effective for improving performance in reverberation is different from what is most effective for ameliorating the effects of degradation caused by spatially-separated interfering sound sources.

R. M. Stern (✉) · A. Menon
Department of Electrical and Computer Engineering
& Language Technologies Institute, Carnegie Mellon University, Pittsburgh,
PA 15213, USA
e-mail: rms@cmu.edu

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_18

1 Introduction

Automatic speech recognition (ASR) is the key technology that enables natural interaction between humans and intelligent machines. Core speech-recognition technology developed over the past several decades in domains such as office dictation and interactive voice-response systems to the point that it is now commonplace for customers to encounter automated speech-based intelligent agents that handle at least the initial part of a user query for airline flight information, technical support, ticketing services, etc. As time goes by, we will come to expect the range of natural human-machine dialog to grow to include seamless and productive interactions in contexts such as humanoid robotic butlers in our living rooms, information kiosks in large and reverberant public spaces, as well as intelligent agents in automobiles while traveling at highway speeds in the presence of multiple sources of noise. Nevertheless, this vision cannot be fulfilled until we are able to overcome the shortcomings of present speech-recognition technology that are observed when speech is recorded at a distance from the speaker.

Two of the major forms of environmental degradation are additive noise of various forms and reverberation. Additive noise arises naturally from interfering speakers, background music, or other sound sources that are present in the environment, and as the signal-to-noise ratio (SNR) decreases, speech recognition becomes more difficult. In addition, the impact of noise on speech-recognition accuracy depends as much on the type of noise source as on the SNR. For example, compensation becomes much more difficult when the noise is highly transient in nature, as is the case with many types of impulsive machine noise on factory floors and gunshots in military environments. Interference by sources such as background music or background speech is especially difficult to handle, as it is both highly transient in nature and easily confused with the desired speech signal. Research directed toward compensating for these problems has been in progress for more than three decades.

Reverberation is also a natural part of virtually all acoustical environments indoors, and it is a factor in many outdoor settings with reflective surfaces as well. The presence of even a relatively small amount of reverberation destroys the temporal structure of speech waveforms. This has a very adverse impact on the recognition accuracy that is obtained from speech systems that are deployed in public spaces, homes, and offices for virtually any application in which the user does not use a head-mounted microphone. It is presently more difficult to ameliorate the effects of common room reverberation than it has been to render speech systems robust to the effects of additive noise, even at fairly low SNRs. Researchers have begun to make meaningful progress on this problem only relatively recently.

In this chapter we discuss some of the ways in which the characteristics of binaural processing have been exploited in recent years to separate and enhance speech signals, and specifically to improve automatic-speech-recognition accuracy in difficult acoustical environments. Like so many aspects of sensory processing, the binaural system offers an existence proof of the possibility of extraordinary performance in sound localization and signal separation, but as of yet we do not know how best to

achieve this level of performance using the engineering tools available in contemporary signal processing.

In the next section we restate very briefly the basic binaural phenomena that have been exploited in contemporary signal enhancement and robustness algorithms for ASR. In Sect. 3, we summarize for the lay person some of the basic principles that underly contemporary ASR systems. We survey a number of computational approaches to improve the accuracy of ASR systems that are motivated by binaural processing in Sect. 4, and we discuss some extensions of these approaches to systems based on deep learning in Sect. 5.

2 Binaural-Hearing Principles

The human binaural system is remarkable in its ability to localize single and multiple sound sources, to separate and segregate signals coming from multiple directions, and to understand speech in noisy and reverberant environments. These capabilities have motivated a great number of studies of binaural physiology and perception. Useful comprehensive reviews of basic binaural perceptual phenomena may be found in a number of sources including Durlach and Colburn (1978), Gilkey and Anderson (1997), Stern et al. (2006), and Kohlrausch et al. (2013), among others, as well as in basic texts on hearing such as Moore (2012) and Yost (2013).

2.1 Selected Binaural Phenomena

While the literature on binaural processing on both the physiological and perceptual sides is vast, the application of binaural processing to ASR is based on a small number of principles:

1. The perceived laterality of sound sources depends on both the interaural time difference (ITD) and interaural intensity difference (IID) of the signals arriving to the two ears, although the relative salience of these cues depends on frequency (e.g., Durlach and Colburn 1978; Domnitz and Colburn 1977; Yost 1981).
2. The auditory system is exquisitely sensitive to small changes of sound, and can discriminate ITDs on the order of 10 μ s and IIDs on the order of 1 dB. Sensitivity to small differences in interaural correlation of broadband noise sources is also quite acute, as a decrease in interaural correlation from 1.00 to 0.96 is readily discernible (e.g., Durlach and Colburn 1978; Domnitz and Colburn 1977). The ITDs arise from differences in path length from a sound source to the two ears, and the IIDs are a consequence of head shadowing, especially at higher frequencies.
3. The vertical position of sounds, as well as front-to-back differentiation in location, is affected by changes in the frequency response of sounds that are imparted by the anatomy of the outer ear, and reinforced by head-motion cues (e.g., Mehrgardt

and Mellert 1977; Wightman and Kistler 1989a, b, 1999). The transfer function from the sound source to the ears is commonly referred to as the *head-related transfer function* (HRTF). HRTFs generally depend on the azimuth and elevation of the source relative to the head, as well as the anatomy of the head and outer ear of the individual.

4. The intelligibility of speech in the presence of background noise or some other interfering signal becomes greater as the spatial separation between the target and masking signals increases. While some of the improvement in intelligibility with greater spatial source separation may be attributed to monaural effects such as a greater effective SNR at one of the two ears, binaural interaction also appears to play a significant role (e.g., Zurek 1993; Hawley et al. 1999).
5. The auditory localization mechanisms typically pay greater attention to the first component that arrives (which presumably comes directly from the sound source) at the expense of later-arriving components (which presumably are reflected off the room and/or objects in it). This phenomenon is referred to as the *precedence effect* or the *law of the first wavefront* (e.g., Wallach et al. 1949; Blauert 1997; Litovsky et al. 1999).

2.2 Models of Binaural Interaction

A number of models have been developed that attempt to identify and explain the mechanisms that mediate the many interesting binaural phenomena that have been observed. For the most part, the original goals of these models had been to describe and predict binaural lateralization or localization, discrimination, and detection data, rather than to improve ASR recognition accuracy. These models are typically evaluated on their ability to describe and predict the perceptual data, the generality of their predictions, and the inherent plausibility of the models in terms of what is known about the relevant physiology. Useful reviews of binaural models may be found in Colburn and Durlach (1978), Stern and Trahiotis (1995, 1996), Trahiotis et al. (2005), Braasch (2005), Colburn and Kulkarni (2005), and Dietz et al. (2017), among other sources.

Most theories of binaural interaction (at least for signals that are presented through headphones) include a model that describes the peripheral response to sound at the level of the fibers of the auditory nerve, a mechanism for extracting ITDs, a mechanism for extracting IIDs, a method for combining the ITDs and IIDs, and a mechanism for developing predictions of lateral position from the combined representation. Models that describe sound localization in the free field typically incorporate information from HRTFs.

Models of Auditory-Nerve Activity

Models of the response to the sounds at the auditory-nerve level typically include (1) a bandpass frequency response, with a characteristic frequency (CF) that provides the greatest response, (2) some sort of half-wave rectification that converts the output of

the bandpass linear filters to a strictly positive number that represents rate of response, and (3) synchrony or “phase locking” in the response to the fine structure of low-frequency inputs and to the envelopes of higher-frequency inputs. Some auditory-nerve models also include (4) enhanced response at the temporal onset of the input and (less frequently) (5) an explicit mechanism for lateral suppression in fibers with a given CF to signal components at adjacent frequencies. These models of auditory-nerve activity can be as simple as the cascade of a bank of bandpass filters, half-wave rectification, and lowpass filtering; more complex and physiologically-accurate models are described in Zhang et al. (2001) and Zilany et al. (2009), among other sources.

Cross-Correlation-Based Models

Most models of binaural interaction include some form of Jeffress’s (1948) description of a neural “place” mechanism as the basis for the extraction of interaural timing information. Specifically, Jeffress postulated a mechanism that consisted of a number of central neural units that recorded coincidences in neural firings from two peripheral auditory-nerve fibers, one from each ear, with the same CF. It was further postulated that the neural signal coming from one of the two fibers is delayed by a small amount that is fixed for a given fiber pair. Because of the synchrony in the response of low-frequency auditory-nerve fibers to low-frequency signals, a given binaural coincidence-counting unit at a particular frequency will produce maximal output when the external stimulus ITD at that frequency is exactly compensated for by the internal delay of the fiber pair. Hence, the external ITD of a simple stimulus could be inferred by determining the internal delay that has the greatest response over a range of frequencies. Colburn (1969, 1973) reformulated Jeffress’s hypothesis quantitatively using a relatively simple model of the auditory-nerve response to sound as Poisson processes, and a “binaural displayer” consisting of a matrix of coincidence-counting units of the type postulated by Jeffress. These units are specified by the CF of the auditory-nerve fibers that they receive input from as well as their intrinsic internal delay. The overall response of an ensemble of such units as a function of internal delay is similar to the running interaural cross-correlation of the signals to the two ears, after the peripheral cochlear analysis (e.g., Stern and Trahiotis 1995). This general representation has been used in a number of computational models of binaural processing for speech recognition, with sound-source locations identified by peaks of the interaural cross-correlation functions along the internal-delay axis.

Figure 1 illustrates how the Jeffress-Colburn mechanism can be used to localize two signals according to ITD. The upper two panels of the figure show the magnitude spectra in decibels of the vowels /AH/ and /IH/ spoken by a male and a female speaker, respectively. The lower panel shows the relative response of the binaural coincidence-counting units when these two vowels are presented simultaneously with ITDs of 0 and -0.5 ms, respectively. The 700-Hz first formant of the vowel /AH/ is clearly visible at the 0-ms internal delay, and the 300-Hz first formant of the vowel /IH/ is seen at the delay of -0.5 ms.

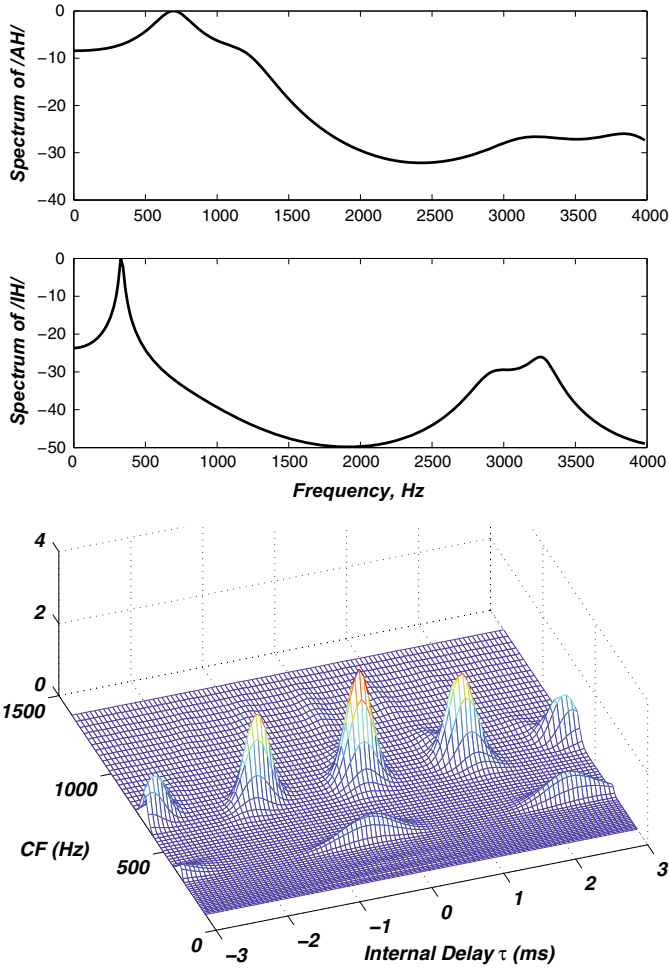


Fig. 1 Upper and central panels: spectrum of the vowels /AH/ and /IH/ as recorded by a male and female speaker, respectively. Lower panel: response of an implementation of the Jeffress-Colburn model to the simultaneous presentation of the /AH/ presented with a 0-ms ITD and the /IH/ presented with a -0.5 ms ITD

It should be noted that the interaural cross-correlation function does not describe IIDs unambiguously, so some additional mechanism must be employed to represent the contributions of IID. For example, Stern and Colburn (1978) multiplied the cross-correlation-based representation of ITD described above by a pulse-shaped function with a location along the internal-delay axis that depends on IID. This model, known as the “position-variable model,” predicts lateral position by computing the centroid of the product of these “timing” and “intensity” functions along the internal-delay axis and then integrating this function over characteristic frequency.

Shamma et al. (1989) proposed an alternative implementation of the Jeffress model, called *stereausis* in which the internal delays are obtained implicitly by comparing inputs of auditory-nerve fibers with slightly mismatched characteristic frequencies, as previously suggested by Schroeder (1977).

Blauert and his colleagues proposed a similar representation (Blauert and Cobben 1978; Blauert 1980). This work was subsequently extended by Lindemann (1986a), who added a mechanism that (among other things) inhibits outputs of the coincidence counters when there is activity produced by coincidence counters at adjacent internal delays. This contralateral inhibition mechanism enables the Lindemann model to describe several interesting phenomena related to the precedence effect (Lindemann 1986b). Gaik (1993) extended the Lindemann mechanism further by adding a second weighting to the coincidence-counter outputs that reinforces naturally-occurring combinations of ITD and IID.

The Equalization-Cancellation Model

The Equalization-Cancellation (EC) model of Durlach and colleagues (e.g., Durlach 1963, 1972) is an additional important alternate model. The EC model was initially formulated to account for binaural detection phenomena, although it has been applied to other psychoacoustical tasks as well (Colburn and Durlach 1978). The model assumes that time-delay and amplitude-shift transformations are applied to the incoming signal on one side in order to *equalize* the masker components of the signals to the two ears. The masker-equalized signals are then subtracted from one another to *cancel* the masker components, leaving the target easily detectable. Stochastic “jitter factors” are applied to the time and amplitude transformations, which limits the completeness of the equalization and cancellation operations, in a fashion that is fitted to the observed limits of human detection performance. The EC model remains popular because of its simplicity and its ability to describe many phenomena. It has been the inspiration for subsequent models (e.g., Breebaart et al. 2001a, b, c), and has also been applied to speech recognition, as will be discussed below.

Detection of Target Presence Using Interaural Correlation

Many phenomena, especially in the area of binaural detection, can be interpreted easily by considering the change in interaural correlation that occurs when a target is added to the masker. The use of interaural correlation was formalized in one binaural early model (Osman 1971) and has been the focus of many experimental and theoretical studies since that time, as reviewed by Trahiotis et al. (2005) among other sources. While cross-correlation-based models that represent ITD, the EC model, and correlation-based models differ in surface structure, it has been shown that under many circumstances they function similarly for practical purposes (e.g., Colburn and Durlach 1978; Domnitz and Colburn 1976).

3 Selected Robust Speech-Recognition Principles

The field of robust automatic speech recognition is similarly vast, and cannot be dealt with in any depth in a review chapter of this scope. The purpose of this section is to provide some insight into the principles of automatic speech recognition that are needed to appreciate the role that binaural processing can play in reducing error rates.

3.1 Basic Speech-Recognition Principles

Automatic speech recognition is essentially a special class of *pattern classification* algorithms, that guess which of a number of possible “classes” of input is actually present. All pattern classification systems operate on the same basic principles: an initial analysis stage performs a physical measurement (of a sound pressure wave, in our case) and transforms that measurement into a set of *features*, or numbers that are believed to be most indicative of the classification task to be performed. These features are typically a stochastic representation that depends on which input class is present. A second decision-making component develops a hypothesis of which of the possible inputs is most likely, based on the observed values of the features. Figure 2 summarizes the major functional blocks of a generic ASR system with binaural pre-processing for signal enhancement. While Fig. 2 depicts a binaural pre-processing module that passes on to the ASR components a restored speech waveform, some of the algorithms we describe produce a restored set of features directly. We briefly discuss the components of the speech recognition system in this section and defer our discussion of the numerous approaches to signal and feature enhancement based on binaural processing to Sect. 4 below.

Feature Extraction

Features for pattern classification systems are generally selected with the goals of being useful in distinguishing the classes to be identified, easy to compute, and not very demanding in storage. With some exceptions, most speech recognition systems

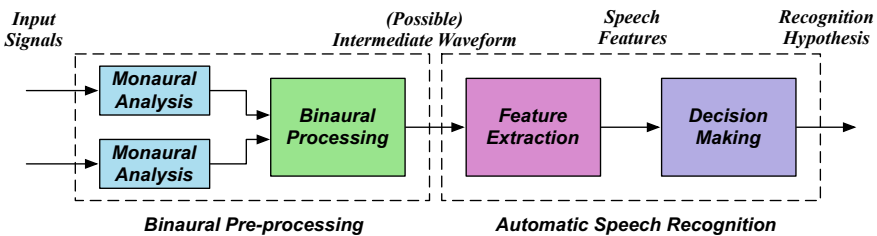


Fig. 2 Basic functional elements of a speech recognition system that includes binaural enhancement

today extract features by first computing the *short-time Fourier transform* (STFT) of the input signal (Allen and Rabiner 1977), typically windowing the incoming signal by a succession of Hamming windows of duration approximately 25 ms, separated by approximately 10 ms. A function related to the log of the magnitude of the spectrum or its inverse transform, the *real cepstrum*, is subsequently computed in each of these analysis frames. In principle, cepstral coefficients are useful because they are nearly statistically independent of one another, and only a small number of them (about 12) are needed to characterize the envelope of the spectrum in each analysis frame. In addition, the cepstral representation separates the effects of the vocal-tract filter (which were believed to be most useful in the early days of the speech recognition) from the effects of the periodic excitation produced by the vocal cords (which had been believed not to be useful at that time).

The most common representations used for feature vectors today are all motivated by crude models of auditory processing. The earliest such representation, *mel frequency cepstral coefficients* (MFCC features, Davis and Mermelstein 1980), multiply the energy spectrum extracted from each analysis frame by a series of triangularly-shaped weighting functions with vertices spaced according to the Mel frequency scale (Stevens et al. 1937) and then summing the product over frequency within each weighting function. With 16-kHz sampling, about 40 Mel weighting functions are typically used. The MFCC coefficients are obtained by computing the inverse discrete cosine transform (DCT) of the summed products. A second set of popular features are extracted using a process known as *perceptual linear prediction* (PLP features, Hermansky 1990), which is based on a more detailed and accurate model of the peripheral auditory system. A more recently-developed third set of features, *power-normalized cepstral coefficients* (PNCC features, Kim and Stern 2016) are more robust to certain types of additive noise and reverberation.

The MFCC, PLP, or PNCC features are typically augmented by additional features that represent the instantaneous power in each analysis frame, as “delta” and “delta-delta” features that serve to represent crudely the first and second derivatives in the power spectrum over time. The delta features are obtained by computing the difference between cepstral coefficients in frames after and before the nominal analysis frame, and the delta-delta features are obtained by repeating this operation. Finally, static effects of linear filtering to the signal are removed by applying either *cepstral mean normalization* (CMN), or *relative spectral analysis* (RASTA) processing (Hermansky and Morgan 1994). CMN subtracts the mean of the cepstral coefficients from each cepstral vector on a sentence-by-sentence basis while RASTA processing passes the cepstral coefficients through a bandpass filter. Both RASTA and CMN serve to emphasize temporal change in the cepstral coefficients and suppress slow drift in their values over time.

Traditional HMM-GMM Decoding

The technologies for determining the most likely word sequence from a spoken utterance have evolved greatly over the decades, and this section will discuss only the most basic elements of speech recognition. From the early 1980s until very recently the dominant speech recognition technology has been the *hidden Markov model*

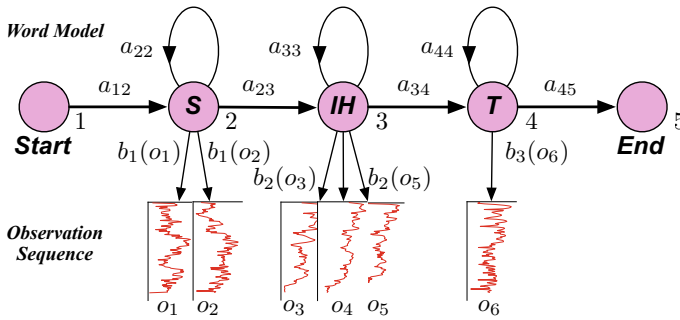


Fig. 3 The hidden Markov model for speech recognition for the word “sit.” See text for details

(HMM, e.g., Rabiner 1989; Rabiner and Juang 1993), and practical systems based on HMMs remain in widespread use today. The HMM representation characterizes the incoming speech waveform as a doubly stochastic process, as depicted in schematic form in Fig. 3. First, the sequence of phonemes that are produced is characterized as a set of five unobserved Markov states which presumably represent the various configurations that the speech production mechanisms may take on and hence the phonemes that are produced. As is the case for all Markov models, the transition probabilities depend only on the current state that is being occupied. Each state transition causes a feature vector to be emitted that is observable, with the probability density of the components of the feature vector depending on the identity of the state transition. Spectra representing a sequence of six observations are shown in the figure. The task of the decoder is to infer the identity of the unobserved state transitions (and hence the sequence of phonemes) from the observed values of the features.

The technologies for implementing this model efficiently and accurately have evolved greatly over decades, and a detailed description is well beyond the scope of this chapter. Briefly, implementing an HMM requires determining the probabilities of the observations given the model parameters, choosing the most likely state sequence given the observations, and determining the model parameters that maximizes the observation probabilities. Details of how to accomplish these tasks are described in standard texts such as Rabiner and Juang (1993) and Gold et al. (2011), as well as in many technical papers. It has been found that the performance of the system depends more critically on the accuracy of the phonetic model (i.e., the probability density function that describes the feature values given the state transitions) than on the probabilities that characterize the state transitions. Gaussian mixture densities are currently the form that is most commonly used for the phonetic models, in part because the parameters of these densities can be estimated efficiently, typically using a form of the expectation-maximization (EM) algorithm (e.g., Dempster et al. 1977). HMMs using Gaussian mixtures for the phonetic models are frequently referred to as “HMM-GMM” systems.

Speech Recognition Using Deep Learning

While the HMM-GMM paradigm has been the dominant speech recognition technology from the early 1980s through the mid-2000s, new approaches to speech recognition based on *deep learning* are becoming more popular. The structures that implement deep learning are frequently referred to as *artificial neural networks* or *computational neural networks*. The general organization and function of computational neural networks was originally motivated by basic neural anatomy and physiology, although the classifiers have evolved considerably over the years without any necessary tie to neural processing by living beings.

While the basic approaches to pattern classification using computational neural networks have been known for some time (e.g., Rosenblatt 1959; Lippmann 1987, 1989; Bourlard and Morgan 1994), these approaches have become more effective and practical in recent years because of a better understanding of the capabilities of the underlying mathematics, the widespread availability of much larger databases for training, and much faster computing infrastructure, including the availability of *graphics processing units* (GPUs), which are particularly well suited for many of the core computations associated with neural networks.

Figure 4 is a crude depiction of the simplest type of deep neural network (DNN) known as the *multilayer perceptron* (MLP). The system consists of an input layer of units, one or more “hidden layers,” and an output layer. Typically the units in a given layer are a weighted linear combination of the values of the units of the previous layer, with the values of the weights trained to minimize the mean square error of the result, using a technique based on gradient descent known as *back propagation* (e.g., Haykin 2018). In many cases DNN classifiers make use of observed values of multiple feature sets (e.g., Mitra et al. 2017). In general, computational neural networks have the advantage of being able to model probability density functions of any form by learning their shape by observing large numbers of training examples. They have the disadvantage of requiring more training data than conventional HMM-GMM systems, and they may not generalize as well as HMM-GMM-based systems.

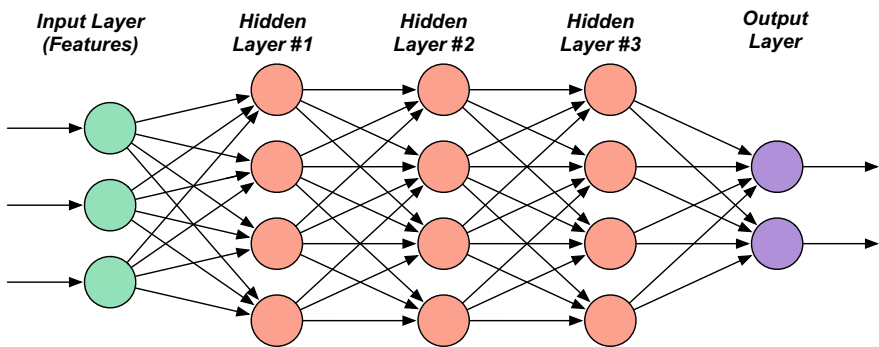


Fig. 4 Standard structure of a feedforward MLP. The network is considered to be “deep” if there are two or more hidden layers

While neural networks were initially used to produce better phonetic models in a system that incorporated a traditional HMM for the decoding component (e.g., Hermansky et al. 2000), other architectures are becoming more popular in which the entire end-to-end speech recognition process is performed using a chain of deep neural networks (e.g., Miao and Metze 2017). Nevertheless, they are increasingly popular because they provide consistently better acoustic-phonetic models than the traditional Gaussian mixtures. The technologies of deep learning have undergone explosive growth and development in recent years, and the reader is referred to standard texts and tutorials such as Goodfellow et al. (2016) and Nielsen (2016) for detailed explanations of the technology.

3.2 *Signal Processing for Improved Robustness in ASR*

We discuss briefly in this section some of the traditional approaches that have been applied to signals to improve recognition accuracy in ASR systems. This field is vast, and has been the object of very active research for decades. Excellent recent reviews of a variety of techniques may be found in Virtanen et al. (2012). In this section we focus on basic feature enhancement techniques, missing-feature approaches, and the uses of multiple microphones.

Feature-Based Compensation for Noise and Filtering

Many successful approaches to robustness in ASR are direct descendants of approaches that were first proposed to enhance speech for human listeners. For example, *spectral subtraction* (Boll 1979), reduces the effects of additive noise by estimating the magnitude of the noise spectrum and subtracting it on a frame-by-frame basis from the spectrum of the signal, reconstructing the time-domain signal with the original unmodified phase. This approach was the basis of dozens if not hundreds of subsequent noise-mitigation algorithms. Stockham et al. (1975) proposed the use of *homomorphic deconvolution* to mitigate the effects of linear filtering by, in effect, subtracting the log magnitude spectrum (or its inverse transform, the real cepstrum) of an estimate of the sample response of the unknown linear filter. A simplified version of this approach is the basis for the cepstral mean normalization that is widely used in ASR systems today.

Joint compensation for the effects of noise and filtering is complicated by the fact that they combine nonlinearly: noise is additive in the time and frequency domains while the effects of filtering are additive in the log spectral and cepstral domains. One particularly successful approach has been the *vector Taylor series* (VTS) algorithm (Moreno et al. 1996), which models the degraded speech as clean speech passed through an unknown linear filter and subjected to unknown additive noise. The algorithm estimates the parameter values that characterize the filtering and noise in a fashion that maximizes the probability of the observations. A recent review of VTS and a number of other techniques motivated by it may be found in Droppo (2013). Algorithms like VTS can provide good improvements to recognition accuracy when

the statistics characterizing the noise and filtering are quasi-stationary while parameters are being estimated, but they are less effective when disturbances are more transitory as in the case of background music or a single interfering speaker. The use of missing-feature approaches as described below has been more effective for these signals.

Computational Auditory Scene Analysis and Missing-Feature Approaches

Modern missing-feature approaches to robust recognition are inspired by Bregman's seminal work (Bregman 1990) in *auditory scene analysis*. Bregman examined the cues that people appear to use in order to segregate and cluster the various components that belong to individual sound sources while perceiving multiple sources that are presented simultaneously. Cues that have proved to be useful include commonalities in onset, amplitude modulation, frequency modulation, and source location, along with harmonicity of components, among others.

Computational auditory scene analysis (CASA) refers to a number of approaches that attempt to emulate the perceptual segregation of sound sources using computational techniques (e.g., Brown and Cooke 1994; Cooke and Ellis 2001; Wang and Brown 2006). The implementation of CASA to isolate the desired signal for an ASR system typically begins by determining which components of the incoming signal are dominated by the target signal and hence not distorted or "missing." In ASR systems, the initial representation is typically in the form of a spectro-temporal display such as a spectrogram. Consideration of only those elements that are relevant or undistorted can be thought of as a multiplication of the components of the spectrogram by a "binary mask" (if "yes-no" decisions are made concerning the validity of a particular spectro-temporal component) or by a "ratio mask" (if probabilistic decisions are made). Once a mask is developed, speech recognition is performed by considering only the subset of components that are considered to be "present" (e.g., Cooke et al. 2001), or by inferring the values of the "missing" features (e.g., Raj et al. 2004) and performing recognition using the reconstructed feature set.

While signal separation and subsequent ASR using CASA techniques can be quite effective if the binary or ratio mask is estimated correctly (e.g., Cooke et al. 2001; Raj et al. 2004; Raj and Stern 2005), estimating the mask correctly is frequently quite difficult in practice, especially when little is known a priori about the nature of the target speech and the various sources of degradation. One singular exception to this difficulty in estimating the masks correctly arises when signals are separated in space and the target location is known, as components can be relatively easily separated using ITD-based and IID-based information. For this reason, separation strategies motivated by binaural hearing have been quite popular over the years for speech recognition systems that make use of two microphones.

Figure 5 shows sample spectrograms of signals separated according to ITD in anechoic and reverberant rooms using two microphones. The speech sources were placed 2 m from the microphones, and at an angle of $\pm 30^\circ$ from the perpendicular bisector of a line connecting the microphones. The microphones were 4 cm apart and the room impulse response (RIR) simulation package McGovern (2004) was

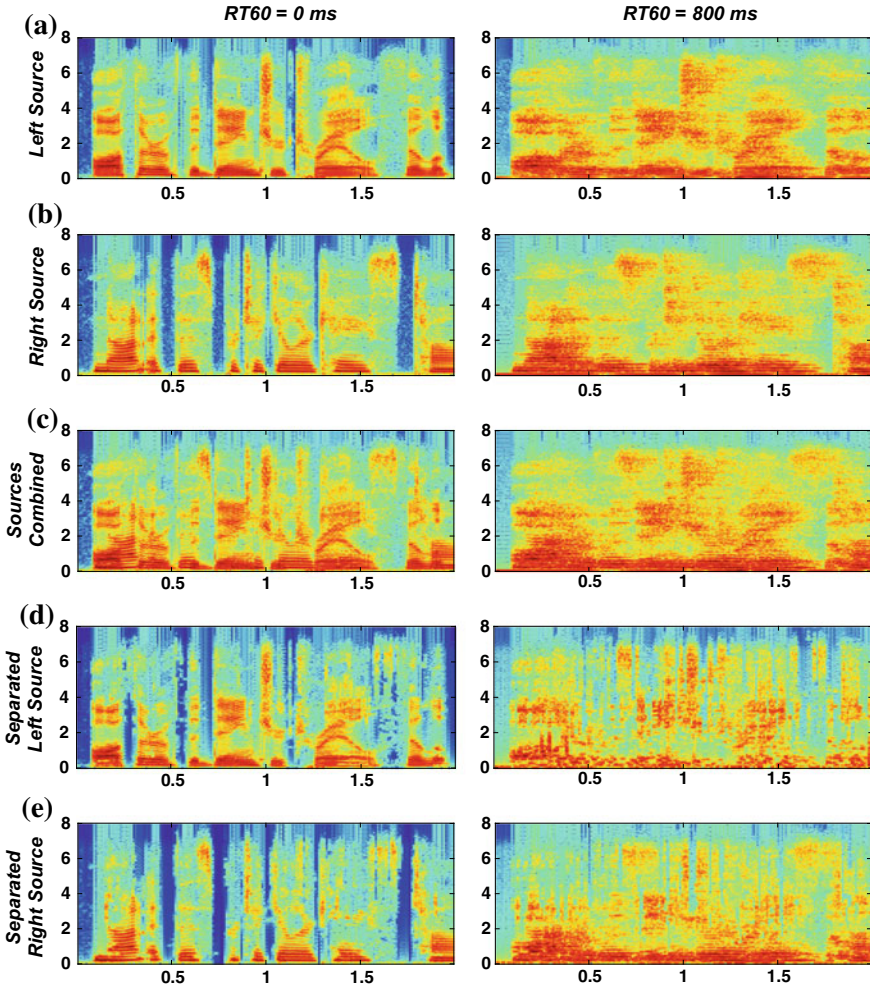


Fig. 5 Sample spectrograms of two speech signals separated according to ITD. The original signals in the left column are clean speech, while the signals in the right column were convolved with a simulated room impulse response with a reverberation time of 800 ms. The spectrograms represent **a** the signal on the left side, **b** the signal on the right side, **c** the two signals combined, **d** the signal on the left side separated from the combined signal according to ITD, **e** the separated signal on the right side. The horizontal axis is time in s and the vertical axis is frequency in kHz

used to develop the simulated impulse responses of the room. The rows of the figure depict, in order, spectrograms of the left speech source, the right speech source, the two sources combined, the separated left source, and the separated right source. By comparing the spectrograms in rows (a) and (d), and (b) and (e), it can be seen that the separation is much more effective when the speech is not reverberated.

Conventional Signal Processing Using Multiple Microphones

The benefit provided by any approach that attempts to improve ASR accuracy using binaural approaches must be compared to the improvement produced by a similar configuration of microphones using conventional techniques. These conventional approaches, frequently referred to as *beamforming algorithms*, attempt to develop a response that is most sensitive to signals coming from a particular “look direction” while either being less sensitive to sources from other directions or actively nulling the responses to these other sources. Classical multi-microphone signal processing techniques are highly developed and discussed in texts including Johnson and Dudgeon (1993) and Van Trees (2004). Recent results concerning the application of multi-microphone techniques to ASR are summarized in Kumatani et al. (2012).

The simplest multi-microphone technique is *delay-and-sum beamforming* in which the path length differences from the target source to the various microphones are compensated for by time delays imposed by the system to ensure that the target signal components from the various microphones always arrives at the same time to the system, creating constructive interference. Signal components from other directions will combine constructively or destructively across the microphones, and hence they would be reinforced to a lesser degree, on average. Because the actual directional sensitivity depends on an interaction between the wavelength at a given frequency, the directivity pattern for delay-and-sum beamforming varies with frequency. Generally the width of the main lobe decreases as frequency increases, and eventually “spatial aliasing” will occur when an interfering signal component arrives at a frequency and azimuth such that the distance between the microphones becomes greater than half a wavelength. These frequency effects can be mitigated by the use of nested arrays with different element spacings (e.g., Flanagan et al. 1985) and by the use of *filter-and-sum* beamforming techniques in which the fixed delays in delay-and-sum beamforming are replaced by discrete-time linear filters which can in principle impose different delays at different frequencies for each microphone.

Modern techniques such as the *minimum variance distortionless response* (MVDR) method use minimum-mean-square estimation (MMSE) techniques that seek to maintain a fixed frequency response in the look direction while at the same time suppressing the response from the directions of arrival of the most powerful interfering sources (e.g., Van Trees 2004). The performance of these optimum linear signal processing approaches to multi-microphone beamforming also degrades in reverberant environments because the phase incoherence imposed by the reverberance causes the estimation of important statistics such as the auto- and cross-correlations of the signals across the microphones to become much less accurate. McDonough and others have achieved some success with the use of objective functions based on negative entropy or kurtosis as the basis for optimizing the filter-coefficient values (e.g., Kumatani et al. 2012). These statistics drive the coefficients of the arrays to produce output amplitude histograms that are “heavier” in the tails, which corresponds to output that is more speech-like than the Gaussian densities that characterize sums of multiple noise sources.

4 Binaural Technology in Automatic Speech Recognition

In this section we describe and discuss selected methods by which ASR accuracy can be improved by signal-processing approaches that are motivated by binaural processing. Most of the systems considered improve ASR accuracy by some sort of selective reconstruction of the target signal using CASA-motivated techniques, which use differing approaches to identify the subset of spectro-temporal components in the input that are dominated to the greatest extent by the target signal. The most common approach makes this determination by comparing measured ITDs and IIDs for each spectro-temporal component to the values of these parameters that would be observed from a source arriving from the putative target direction, as described below in Sect. 4.2. A second approach is based on the value of the overall normalized interaural cross-correlation, as spectro-temporal components with high interaural cross-correlation are more likely to be dominated by a single coherent target signal, as described in Sect. 4.4. A third approach implements a modification to the EC model, in which the two inputs are equalized according to the nature of the *target* signal, and then subtracted from one another, as described in Sect. 4.5. This causes the spectro-temporal components that are dominated the most by the target signal to change by the greatest amount.

In addition to the three methods above used to identify the most relevant spectro-temporal components of the input, the systems proposed also differ in other ways including the following.

- The extent to which a particular system is intended to provide a complete auditory scene analysis, including identification, localization, and classification of multiple sources versus simply providing useful enhancement of a degraded primary target signal for improved speech recognition accuracy.
- Whether the location of the desired target is expected to be estimated by the system or is simply assumed to be known a priori.
- Whether a particular system is designed to receive its input from two ears on a human or manikin head rather than two (or more) microphones in the free field. The use of a real or simulated head provides IIDs and the opportunity to use them to disambiguate the information provided by ITD analysis. In contrast, systems that do not include an artificial head are typically easier to implement, and the absence of a head facilitates the use of more than two microphones.
- Whether a particular system works by reconstructing a continuous-time enhanced speech waveform that is processed by the normal front end of an ASR system or whether it simply produces enhanced features representing the input such as cepstral coefficients and inputs these enhanced features directly into the ASR system.
- The nature of the acoustical environment, including the presence or absence of diffuse background noise, coherent interfering sound sources, and/or reverberation, etc. within which a particular system is designed to operate.

It is worth noting that researchers at the University of Sheffield and Ohio State University, working in collaboration or independently, have provided the greatest number of contributions to this field over the years, both in terms of fundamental principles and system development. Interesting contributions over many years have also been provided by groups at the universities at Bochum and Oldenburg in Germany, as well as a number of other locations around the world including our own university.

The representative systems considered do not sort themselves into convenient mutually-exclusive categories, so we somewhat arbitrarily have sorted our discussion according to how the most relevant spectro-temporal target components are identified, as discussed above. We begin with a brief summary of some of the earliest attempts to apply binaural processing to improve ASR accuracy. We then summarize the organization of representative systems based on extraction of ITD and IID information using CASA principles. We conclude with a discussion of the use of onset enhancement to ameliorate the effects of reverberation, the development of systems based on interaural coherence, and approaches based on the EC model.

4.1 *Early Approaches*

Lyon (1984) proposed one of the first systems applying binaural-hearing principles, using a computational model of auditory-nerve activity from two sources as an input to a Jeffress-like network of coincidence-counting units. He suggested that this structure could be applied to multiple applications including ASR. While Lyon's system was not evaluated quantitatively because the ASR systems of the day were mathematically primitive and computationally costly, he noted that this approach appeared to provide a stable spectral representation for vowels as well as source separation according to ITD.

Most evaluations of ASR with binaural processing in the early period consisted of the concatenation of an existing binaural model with a speech recognition system. For example, Bodden (1993) described an early CASA-based system, called the Cocktail-Party-Processor (CPP) that had many of the elements of later systems, implementing a structure suggested by Blauert (1980). The CPP included HRTFs that introduced frequency-dependent ITDs and IIDs based on angle of arrival, a relatively simple auditory-nerve model that included bandpass filtering, half-wave rectification, lowpass filtering, and saturation of the rate of response. Binaural processing in the CPP incorporated the Lindemann (1986a) model with contralateral inhibition, which predicted certain precedence-effect phenomena and appropriate interactions between ITD and IID, and the additional contributions of Gaik (1993), which developed lateralization information from ITDs and IIDs in a fashion that was cognizant of the natural combinations of these interaural differences as observed in HRTFs. In later work, Bodden and Anderson (1995) used a simple speech recognizer, the self-organizing feature map (SOFM) of Kohonen (1989), and demonstrated improved ASR accuracy for simple phonemes in the presence of spatially-separated noise,

especially at lower SNRs. DeSimio et al. (1996) obtained similar results with a different auditory-nerve model (Kates 1991) and Shamma's stereausis model (Shamma et al. 1989) to characterize the binaural interaction.

4.2 Systems Based on Direct Extraction of ITD and IID Information

By far the most common application of binaural principles to ASR is through systems that implement computational auditory scene analysis using direct extraction of ITDs and IIDs in some fashion, as depicted in Fig. 6. In general, these systems attempt to estimate the extent to which each spectro-temporal component of the input is dominated by the target signal based on ITDs and IIDs that are extracted. We summarize in this section a few of the methods that are used to implement each component in representative systems.

Extraction of Natural Interaural Differences

As noted above, a number of the systems develop their hypotheses from naturally-extracted interaural differences (e.g., Roman et al. 2003; Palomäki et al. 2004; Srinivasan et al. 2006; Brown et al. 2006; Harding et al. 2006; May et al. 2011, 2012). While these systems typically made use of measured HRTFs (e.g., Gardner and Martin 1994) obtained through the use of the KEMAR manikin (Burkhard and Sachs 1975), they could also have been obtained in principle using small microphones in the ear canals (e.g., Wightman and Kistler 1989a). Because the relationship between ITD and the azimuth of the source location in HRTFs depends weakly on frequency, some systems (e.g., Roman et al. 2003; Palomäki et al. 2004; May et al. 2011) incorporated an explicit mapping table that converts ITD into putative arrival angle in a manner that is consistent across all frequencies. The IIDs show significant dependencies on both azimuth and frequency. For the most part, the various sound sources were assumed to be at the same elevation as the microphones.

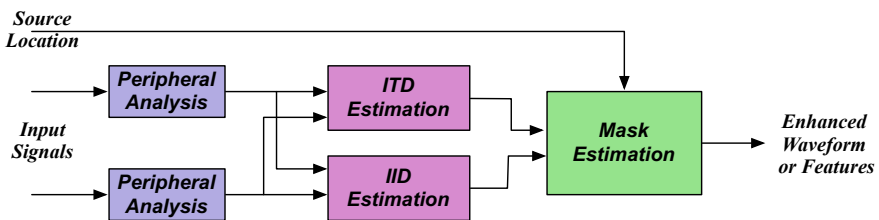


Fig. 6 Functional blocks of a CASA-based system that extracts ITDs and IIDs directly. The input may be from free-field microphones or through a real or simulated head. The source location may be known or estimated

Other systems (e.g., Aarabi and Shi 2004; Park and Stern 2009; Kim et al. 2009) work from free-field input without an artificial head or HRTFs and, consequently, the masks that are produced cannot make use of IID information.

Peripheral Auditory Processing

All binaural processing systems incorporate some abstraction of the frequency-dependent processing imparted by the peripheral auditory system. The most common approach (e.g., Roman et al. 2003; Palomäki et al. 2004; Harding et al. 2006) is to use a bank of 40–128 Gammatone filters (Patterson et al. 1988), followed by half-wave rectification, lowpass filtering (which provides envelope extraction at higher frequencies), and in some cases nonlinear compression of the resulting signal. Other systems (e.g., Kim et al. 2009) simply compute the short-time Fourier transforms (STFTs) of the two input signals, from which ITDs and IIDs can be inferred by comparison of the magnitudes and phases for each spectro-temporal component.

Estimation of ITDs and IIDs

There are multiple ways of extracting ITDs from the results of the peripheral processing. The most common approach is to compute a variant of the *normalized interaural cross-correlation function* at each frequency:

$$R[m, k] = \frac{\sum_{n=0}^{N-1} x_{L,k}[n]x_{R,k}[n-m]}{\sqrt{\sum_{n=0}^{N-1} x_{L,k}^2[n]}\sqrt{\sum_{n=0}^{N-1} x_{R,k}^2[n-m]}}$$

where $R[m, k]$ is the normalized interaural cross-correlation as a function of lag m and frequency index k , and $x_{L,k}[n]$ and $x_{R,k}[n]$ are the left and right signals, respectively, after peripheral processing at frequency k . The *interaural cross-covariance function* is a very similar statistic in which the means are subtracted from $x_{L,k}[n]$ and $x_{R,k}[n]$ before further computation. In both cases, the ITD is typically inferred by searching for the value of m that maximizes $R[m, k]$ in each frequency channel (e.g., Roman et al. 2003; Brown et al. 2006; Harding et al. 2006; May et al. 2011, 2012). Because this maximum may not occur at an integer value of m , polynomial or exponential interpolation is typically performed in the region of the maximum, with the true maximum value determined either analytically or via a grid search. In some systems (e.g., Roman et al. 2003; Palomäki et al. 2004) the cross-correlation function is summed over frequency before the maximum is obtained. This is useful because it reduces ambiguity in identifying the true ITD of a source, particularly for larger ITDs and higher frequencies by emphasizing ITDs that are consistent over frequency as in human auditory processing (Stern et al. 1988). In addition, the cross-correlation function may be “skeletonized” by replacing the normalized cross-correlation function by Gaussians located at the values of m that maximize $R[m, k]$ at each frequency (e.g., Roman et al. 2003; Palomäki et al. 2004). This can be helpful in interpreting the responses to binaural signals that include multiple sound sources.

Systems that use STFTs as the initial stage of processing can infer ITD by calculating the phase of the product of one STFT multiplied by the complex conjugate of

the other (which represents the instantaneous cross-power-spectral density function), and dividing the phase by frequency to convert to ITD (e.g., Aarabi and Shi 2004; Srinivasan et al. 2006; Kim et al. 2009). ITDs can also be estimated by comparing the times at which zero crossings in the signals after peripheral processing appear (Park and Stern 2009).

In contrast, IID estimation is relatively straightforward, and is almost always estimated as the ratio of signal energies, expressed in decibels for each spectro-temporal component of the two inputs.

Mask Estimation

As noted above, the masks developed by the systems are intended to represent the extent to which a given spectro-temporal component is dominated by the target component rather than the various interfering sources or maskers in the input. The target location is either estimated by the system in initial processing (e.g., Roman et al. 2003; Palomäki et al. 2004; May et al. 2011, 2012), or by assuming a location for the target (typically directly to the front of a head or at zero ITD for two microphones). The masks are obtained by evaluating (either explicitly or implicitly) the probability of the observed ITDs and IIDs given the putative location of the sound source. For many systems these probabilities are estimated from training data, although the distributions of ITDs and especially IIDs are affected by the amount of reverberation in the environment. As noted above, the masks are either binary masks (i.e. equal to zero or one for each spectro-temporal component) or ratio masks (which typically take on values equal to a real number between zero and one). Because the peripheral filters are narrowband, the maxima of the interaural cross-correlation function repeat periodically along the lag axis, and the IIDs provide information that is helpful in disambiguating the cross-correlation patterns.

Another much more simple method approach is to compare the ITD estimated for each spectro-temporal component to the ITD associated with the target location, and to assign a value of one to those components that are sufficiently “close” to the target ITD using a binary or probabilistic decision (e.g., Kim et al., 2009).

System Evaluation and Results

Once the mask that identifies the undistorted target components is developed, some systems use bounded marginalization (Cooke et al. 2001) to recognize the target speech based on the components that are most likely to be informative (e.g., Roman et al. 2003; Palomäki et al. 2004; May et al. 2012). Other systems (e.g., Kim et al. 2009, 2010, 2012) reconstruct the waveform from a subset of spectro-temporal components that are deemed to be useful.

The motivations and goals of the systems considered in this subsection vary widely, making it difficult to compare them (along with other similar systems) directly. Nevertheless, a few generalizations can be made:

- Objective speech recognition and speaker identification accuracy obtained follow trends that would normally be expected: recognition accuracy degrades as SNR decreases, as the spatial separation between the target speaker and interfering

sources decreases, and as the amount of reverberation increases. Recognition or identification accuracy is invariably substantially better with binaural processing compared to baseline systems that use only a single microphone.

- In situations where they can be compared directly, the use of ratio masks tends to provide greater recognition accuracy than the use of binary masks. If binary masks are used, we have found in our own work that accuracy is improved when the binary masks are smoothed over time and over frequency. The temporal smoothing can be accomplished by simply averaging the mask values at a given frequency over a few adjacent frames. We have used “channel weighting” to accomplish the frequency smoothing, which is in essence a multiplication of the Gammatone frequency response representing each channel by the corresponding value of the binary masks and summing over frequency (e.g., Kim et al. 2009).
- In the single case where zero-crossing-based ITD extraction was compared to ITDs by searching for the maximum of the interaural cross-correlation function, the zero-crossing approach provided better results (Park and Stern 2009).
- Source localization strategies in systems such as those by Roman et al. (2003), Palomäki et al. (2004), and May et al. (2011) appear to be effective, and their performance with multiple and moving sources should improve over time.

4.3 Robustness to Reverberation Using Onset Emphasis

As noted in Sect. 2.1, many classic psychoacoustical results indicate that the auditory localization mechanism places greater emphasis on the first-arriving components of a binaural signal (e.g., Wallach et al. 1949; Blauert 1997; Litovsky et al. 1999), a phenomenon known as the “precedence effect.” More recent studies (e.g., Stecker et al. 2013) confirm that the lateralization of brief steady-state sounds such as tones and periodic click trains based on ITDs and IIDs appears to be strongly dominated by binaural cues contained in the initial onset portion of the sounds. In addition, Dietz et al. (2013) have shown that the fine-structure ITD in slow sinusoidal amplitude modulation appears to be sampled briefly during the rising-envelope phases of each modulation cycle, and is not accessed continuously over the duration of the sound.

The precedence effect is clearly valuable in maintaining a constant image location in reverberant environments when the instantaneous ITDs and IIDs produced by a sound source are likely to vary with time (Zurek et al. 2004). In addition, Blauert (1983) and others have noted that the precedence effect is likely to play an important role in increasing speech intelligibility in reverberant environments. While precedence has historically been assumed to be a binaural phenomenon (e.g., Lindemann 1986a), it could also be mediated by monaural factors such as an enhancement of the onsets of envelopes of the auditory response to sound on a channel-by-channel basis at each ear (Hartung and Trahiotis 2001).

Motivated by the potential value of onset enhancement for improved recognition accuracy in reverberation, several research groups have developed various methods of enhancing envelope onsets for improved recognition accuracy in reverberant

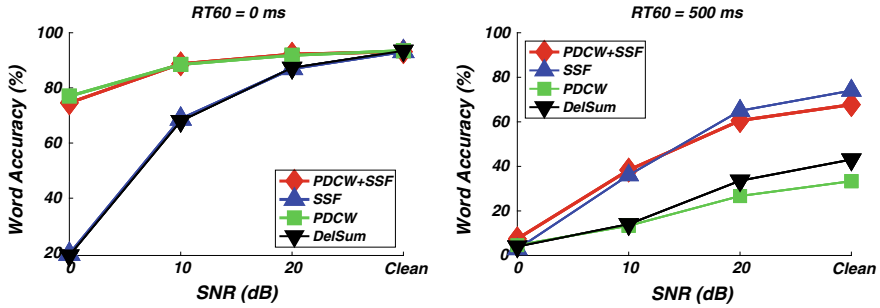


Fig. 7 Comparison of ASR accuracy obtained using the PDCW algorithm (which separates signals according to ITD) and the SSF algorithm (which enhances onsets of signal components) in the presence of additive noise and reverberation, plotted as a function of SNR. See text for further details

environments. Palomäki et al. (2004) described an early comprehensive CASA-based binaural model that included an explicit mechanism for onset enhancement for precedence, along with other components including HRTFs, skeletonization of the cross-correlation representation, and the use of IIDs at higher frequencies as a consistency check on the estimated binary mask. More recent algorithms that incorporate onset enhancement include the algorithm known as Suppression of Slowly-varying components and the Falling edge of the power envelope (SSF) (Kim and Stern 2010), the temporal enhancement component of the STM algorithm (Kim et al. 2011), and the SHARP algorithm (Cho et al. 2016). All of these approaches incorporate nonlinear processing of the energy in the spectral envelopes to enhance transients, and they can be considered to be improved versions of the envelope enhancement approach suggested by Martin (1997) that had been used by Palomäki et al. (2004) and others. The temporal suppression components in Power-Normalized Cepstral Coefficients (PNCC, Kim and Stern 2016) provide similar benefit in reverberation, but to a more limited extent.

Figure 7 compares selected sample recognition accuracies for the DARPA Resource Management (RM1) task using implementations at Carnegie Mellon University of two of the approaches described above. The phase-difference channel weighting algorithm (PDCW) (Kim et al. 2009) improves ASR accuracy by separating the target speech signal from the interfering speaker according to ITD, as in other algorithms discussed in Sect. 4.2. The SSF algorithm (Kim and Stern 2010) improves ASR accuracy by enhancing the onsets and suppressing the steady-state portions of subband components of the incoming signals, as described in this section. The data in Fig. 7 consist of a target signal directly in front of a pair of microphones in the presence of an interfering speech source at an angle of 30° as well as an uncorrelated broadband noise source. The figure plots recognition accuracy obtained using delay-and-sum beamforming, PDCW, SSF, and the combination of PDCW and SSF. Results are plotted as a function of SNR for simulated reverberation times of 0 (left panel) and 500 ms (right panel). We note that the PDCW and SSF algorithms pro-

vide complementary benefits in the presence of noise and reverberation: PDCW is highly effective, even in the presence of substantial noise if there is no reverberation present, but it provides no benefit when substantial reverberation is present in the acoustical environment. SSF, on the other hand, provides substantial benefit in the presence of reverberation for the reverberation depicted, but it is ineffective in the presence of substantial additive noise. Remediation for the effects of noise is more effective than for reverberation, at least for these two algorithms. We note that the results in Fig. 7 were obtained using an HMM-GMM system that was trained on clean speech. Training in multiple acoustical environments and/or using DNNs for decoding reduces the magnitude of the differences of these results.

These results suggest that the choice of which robustness approach is best in a given situation will depend on the spatial separation of target and masker components as well as the degree of reverberation in a given acoustical environment. The combination of SSF and PDCW almost always provides better performance than is observed with either algorithm by itself. While we used data from our own group for convenience in these comparisons, we believe that the use of information from ITDs (and more generally IIDs as well) to provide robustness against spatially-separated interfering sources and the use of onset enhancement to provide robustness against reverberation are generally effective across a wide range of conditions.

Pilot results from our laboratory indicate that better recognition accuracy is obtained when precedence-based onset emphasis is imposed on the input signals monaurally before binaural interaction, rather than after the binaural interaction.

4.4 Robustness to Reverberation Based on Interaural Coherence

A number of researchers have developed methods to enhance a target signal by giving greater weight to spectro-temporal components that are more “coherent” from microphone to microphone. The original motivation for much of this work was the seminal paper by Allen et al. (1977) who proposed that the effects of reverberation can be removed from a signal by performing a subband analysis, compensating for the ITDs observed in each frequency band, and applying a weighting in each frequency channel that is proportional to the normalized cross-correlation observed in each frequency band.

In subsequent work, Faller and Merimaa (2004) proposed that the salience of a spectro-temporal component representing a particular ITD and frequency can be characterized by a running normalized interaural cross-correlation function similar to the equation in Sect. 4.2 but updated using a moving exponential window in running time. The value of this statistic at the lag that produces the maximum interaural cross-correlation can be taken as a measure of the interaural coherence as a function of frequency.

In recent years a number of researchers have developed various models that predict the *coherent-to-diffuse energy ratio* (CDR) or the closely-related *direct-to-reverberant energy ratio* (DRR) in a given environment (e.g., Jeub et al. 2009, 2010, 2011a; Thiergart et al. 2012; Westermann et al. 2013; Zheng et al. 2015). In general, the various authors use a measure similar to that proposed by Faller and Merimaa to estimate the coherent energy of the target speech and a model of the room acoustics to estimate the energy in the reverberant field. The papers differ in the assumptions that they make about the acoustics of the room, and about the geometry of the head. As a representative example we summarize the two-stage processing proposed by Jeub et al. (2010), (2011b) for reducing the impact of reverberation. In the first stage, steering delays are imposed in the input at each frequency to compensate for differences in the path lengths from the desired source to the various microphones, and spectral subtraction is performed to suppress the effects of late reverberation. In the second stage, the residual reverberation is attenuated by a dual-channel Wiener filter derived from the coherence of the reverberant field, considering the effects of head shadowing, with the objective being suppression of the spectro-temporal components for which there is little correlation.

The systems described in the studies cited above had all been evaluated in terms of subjective or objective measures of speech quality rather than speech recognition accuracy.

We recently completed a series of experiments (Menon 2018) in which we compared the speech recognition accuracy obtained using a local implementation of the second stage of the algorithm of Jeub et al. (2011b), which enhances binaural signals based on their CDR, to the performance of the SSF algorithm described in Sect. 4.3. We refer to our implementation of spectro-temporal weighting based on CDR as CDRW (for Coherent-to-Diffuse-Ratio Weighting). Some of these results are summarized in Fig. 8, which uses similar signal sets and processing as in the data depicted in Fig. 7, except that the acoustical models used to train the ASR system

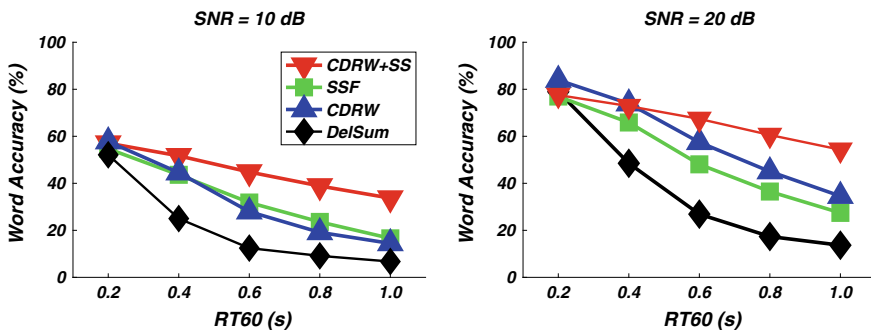


Fig. 8 Comparison of ASR accuracy obtained using the CDRW algorithm (which emphasizes signal components based on their interaural coherence) and the SSF algorithm (which enhances onsets of signal components) in the presence of additive noise and reverberation, plotted as a function of SNR. Signal-to-noise ratios are 10 dB (left panel) and 20 dB (right panel). See text for further details

were obtained using deep neural networks (DNNs). We note that both the CDRW and SSF algorithms are individually effective in reducing error rate in reverberant environments in the presence of an interfering speaker, and that SSF becomes more useful as the reverberation time increases. Moreover, the impacts of the two approaches are complementary in that best results are obtained when the two algorithms are used in combination.

4.5 *EC-Based Processing*

The Equalization-Cancellation (EC) model of Durlach and colleagues (e.g., Durlach 1963, 1972) was summarized briefly in Sect. 2.2. In developing predictions for binaural masking experiments, the EC model typically assumes that the auditory system attempts to “equalize” the masker components to the two ears by inserting ITDs and IIDs that compensate for the corresponding interaural differences that are present in the signals, and then “cancel” the masker by subtracting the signals to the two ears after equalization, leaving the target more detectable. Various investigators have proposed extensions to EC processing to accommodate the rapidly-varying fluctuations in overall ITD and IID imposed by speech-like maskers and have demonstrated that this type of processing can predict speech intelligibility (e.g., Beutelmann and Brand 2006; Beutelmann et al. 2010; Wan et al. 2010, 2014).

The current applications of EC-based processing to improve speech recognition accuracy differ from the traditional application of the EC model to predict binaural detection thresholds in that the equalization and cancellation operations are applied to the *target signal* rather than the masking components. This is sensible both because the SNRs tend to be greater in ASR applications than in speech threshold measurements, and because in practical applications there tends to be more useful information available a priori about the nature of the target speech than about the nature of the background noise and interfering signals. In the first application of this approach, Roman et al. (2006) employed an adaptive filter to cancel the dominant correlated signals in the two microphones, which are presumed to represent the coherent target signal. A binary mask is developed by selecting those spectro-temporal components that are the most affected by the cancellation, which are presumed to be the components most dominated by the target speech. This approach provided better ASR results for reverberated speech in the presence of multiple maskers than several other types of fixed and adaptive beamformers. Brown and Palomäki (2011) described a more sophisticated system that determines the ITD that provides maximum signal cancellation, cancels the signals to the two mics using that ITD, and again uses the absolute difference between the cancelled and uncanceled signal as an indicator of the extent to which a spectro-temporal component is dominated by the target speech. A complete signal was reconstructed using an ASR system based on bounded marginalization of the features (Cooke et al. 2001), although only SRT results were provided in the paper. Mi and Colburn (2016), Mi et al. (2017), and Cantu (2018), among others, have developed more recent systems that enhance speech intelligibility in the presence of interfering sources based on EC principles.

5 Binaural Processing Using Deep Learning

As we noted in Sect. 3.1, systems based on deep learning techniques are rapidly superseding conventional HMM-GMM ASR systems over the last decade, in part because of the superior ability of deep learning approaches to develop more general acoustic models. Most of the major current techniques that enable ASR using DNNs are reviewed in Hinton et al. (2012) as well as in the more recent book edited by Watanabe et al. (2017), among other resources.

Similarly, there has been great interest in the use of deep neural networks (DNNs) to perform the classifications needed to develop the binary or ratio masks to enable signal separation based on CASA principles. These approaches are reviewed comprehensively by Wang and Chen (2018), which considers (among other things) the type of mask to be employed, the choice of “training target” that is optimized in the process of training the mask classifier, the input features, the structure of the DNN used for the separation, and the methods by which the signals are separated and subsequently reconstructed.

The first system to use DNNs to separate binaural signals based on interaural differences was described by Jiang et al. (2014) and has components that are found in a number of similar systems. The system includes HRTFs from the KEMAR manikin, and gammatone-frequency cepstral coefficient features (GFCCs, Shao and Wang 2008), which include 64 gammatone filters whose outputs are half-wave rectified and passed through square-root compression. ITDs are estimated using both a complete representation of the normalized cross-covariance function and a single number indicating the estimated correlation lag with maximum magnitude; IID is estimated from the subband energy ratios. Monaural GFCC features were also employed in the mask classification. A mask classifier was developed for each subband, using DNNs with two hidden layers. To avoid convergence and generalization issues with MLPs, the system was pre-trained using a restricted Boltzmann machine (RBM) (Wang and Wang 2013). The performance of this DNN-based CASA system was compared to that of the contemporary source-separation systems DUET (Rickard 2007) and MESSL (Mandel et al. 2010), as well as systems proposed by Roman et al. (2003) and Woodruff and Wang (2013). Compared to the other systems considered, The DNN-based CASA system of Jiang et al. (2014) was found to produce substantially better approximations to the ideal binary mask that would separate the sources correctly. This system also provided improved output SNR in speech enhancement tasks. The use of the full normalized cross-correlation function (as opposed to a single numerical estimate of ITD), and with the direct inclusion of monaural features into the mask-classification process, were found to be valuable contributors to best performance. The system maintained good accuracy, and generalized to test conditions that were not included in the training for a variety of types of interfering sources and reverberant environments.

Other approaches using DNNs have been suggested as well. For example, Araki et al. (2015) have described the use of a denoising auto-encoder (DAE), which is trained to convert a degraded representation of a speech signal into a clean version

of it. The DAE is typically structured in a “bottleneck” configuration, with at least one hidden layer that is smaller in dimensionality than the input and output layers. Estimation of a ratio mask was based on information at each frequency that included IID, ITD (as estimated from phase differences from the two inputs), and an enhanced signal was reconstructed by filtering the input using the mask that was learned by the DAE. Lowest error rates for keyword recognition in the PASCAL CHiME Speech Separation Challenge were obtained when the DNN was trained using a combination of monaural information and a location-based mask, although IID information was not useful in this particular study. Fan et al. (2016) described a similar system that uses a DNN with RBM-based pre-training to develop a binary mask using features that represented monaural information and IID. They observed better enhanced speech intelligibility when IIDs were extracted on a subband basis, but this system did not make use of ITD information.

Two more sophisticated binaural-based systems that separate speech using DNNs were described by Yu et al. (2016) and by Zhang and Wang (2017). The system of Yu et al. estimated ITD and IID by comparing the magnitudes and phases of the STFT components from the two microphones, along with “mixing vectors” that are obtained by combining the two monaural STFT values for each spectro-temporal component. The DNNs used to estimate the mask were in the form of sparse autoencoders, which were initially trained in unsupervised fashion and later stacked to estimate the probability that each component belongs to one of several possible source directions. The system of Zhang and Wang uses both spectral and spatial features, with the spectral features obtained from the output of an MVDR beamformer with a known target location. The spatial features include ITDs represented by the complete normalized cross-correlation function along with its estimated maximum and IIDs calculated energy ratios in each frequency band. The two systems provided dramatic improvements in SNR and/or speech intelligibility for speech enhancement tasks.

The representative examples above provide merely a superficial characterization of the ever-growing body of work devoted to the development of CASA systems using DNNs that are motivated by binaural processing to improve speech-recognition accuracy. It is clear that the use of DNNs to develop the masks for speech separation systems can provide sharply improved performance compared to conventional classification techniques. This is particularly valuable because determining the spectro-temporal components of a complex input that most clearly represent the target is known to be extremely difficult, even using binaural ITD and IID information. The use of DNNs to segregate and enhance the desired target also provides impressive improvements to source-localization accuracy, and to speech intelligibility, both for normal-hearing and hearing-impaired listeners. Nevertheless, this area of research is still in its infancy. For example, there is not yet a clear sense of what type of DNN architecture is best suited for mask estimation, nor is there yet a clear understanding of which monaural and binaural features are the best inputs to the DNN. Furthermore, most of the systems developed have been evaluated only in terms of measures of speech intelligibility or statistics for speech enhancement such as putative improvement in SNR. So far there have been relatively few applications of these approaches to objective tasks such as speech recognition or speaker verification. Assuming that

the most effective ASR or verification systems use a DNN recognizer, it is not yet clear what is the best architecture for the purpose, nor the extent to which the form of the recognizer should be modified to accommodate missing-feature input, nor the extent to which the complete mask-estimation/recognizer-system architecture could be made more efficient or more effective by merging the two systems.

6 Summary

We have described a number of methods by which the principles of binaural processing can be exploited to provide substantial improvements in automatic speech recognition accuracy, particularly when the target speech and interfering sources are spatially separated and the degree of reverberation is moderate. In general, most of these approaches implement aspects of computational auditory scene analysis, using one of four different approaches to determine the mask which identifies the spectrotemporal components that are believed to be dominated by the target signal: direct extraction of ITDs and IIDs, onset emphasis for reverberation, exploitation of the coherent-to-diffuse ratio or related statistics, and exploitation of principles based on the EC model. This is a particularly exciting time to be working in the application of binaural technology to automatic speech recognition because our rapidly-advancing understanding of how to develop classification techniques based on the principles of deep learning is likely to enable the realization of systems that serve their users increasingly effectively in cluttered and reverberant acoustical environments.

Acknowledgements Preparation of this manuscript was partially supported by grants from Honeywell, Google, and Afeka University. A. Menon has been supported by the Prabhu and Poonam Goel Graduate Fellowship Fund and the Jack and Mildred Bowers Scholarship in Engineering. R. Stern is deeply grateful to the many mentors, colleagues, and friends in the binaural-hearing and speech-recognition communities that have informed this analysis, including especially H. S. Colburn, C. Trahiotis, B. Raj, and R. Singh. The authors also thank E. Gouvêa, C. Kim, A. Moghimi, H.-M. Park, and T. M. Sullivan for many experimental contributions and general insight into these phenomena. Thanks are further due to two anonymous reviewers for valuable comments and suggestions.

References

- Aarabi, P., and G. Shi. 2004. Phase-based dual-microphone robust speech enhancement. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 34: 1763–1773.
- Allen, J.B., D.A. Berkley, and J. Blauert. 1977. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *Journal of the Acoustical Society of America* 62 (4): 912–915.
- Allen, J.B., and L.R. Rabiner. 1977. A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE* 65 (11): 1558–1564.

- Araki, S., T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani. 2015. Exploring multi-channel features for denoising-autoencoder-based speech enhancement. In *Proceedings on IEEE International Conference on Acoustics, Speech and Signal Processing*, 116–120.
- Beutelmann, R., and T. Brand. 2006. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *Journal of Acoustical Society of America* 120: 331–342.
- Beutelmann, R., T. Brand, and B. Kollmeier. 2010. Revision, extension, and evaluation of a binaural speech intelligibility model. *Journal of Acoustical Society of America* 127: 2479–2497.
- Blauert, J. 1980. Modeling of interaural time and intensity difference discrimination. In *Psychophysical, Physiological, and Behavioural Studies in Hearing*, eds. G. van den Brink, and F. Bilsen, 412–424. Delft: Delft University Press.
- Blauert, J. 1983. Review paper: Psychoacoustic binaural phenomena. In *Hearing—Physiological Bases and Psychophysics*, eds. R. Klinke, and R. Hartmann, 182–189. Heidelberg: Springer-Verlag.
- Blauert, J. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*, 2nd ed. Cambridge, MA: MIT Press.
- Blauert, J., and W. Cobben. 1978. Some considerations of binaural cross-correlation analysis. *Acustica* 39: 96–103.
- Bodden, M. 1993. Modelling human sound-source localization and the cocktail party effect. *Acta Acustica* 1: 43–55.
- Bodden, M., and Anderson, T.R. 1995. A binaural selectivity model for speech recognition. In *Proceedings of Eurospeech 1995* (European Speech Communication Association).
- Boll, S.F. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27(2), 113–120.
- Bourlard, H., and Morgan, N. 1994. *Connectionist Speech Recognition: A hybrid approach*. Kluwer Academic Publishers.
- Braasch, J. 2005. Modelling of binaural hearing. In *Communication Acoustics*, ed. J. Blauert, Chap. 4, 75–108. Berlin: Springer-Verlag
- Breebaart, J., S. van de Par, and A. Kohlrausch. 2001a. Binaural processing model based on contralateral inhibition. I. Model structure. *Journal of the Acoustical Society of America* 110: 1074–1088.
- Breebaart, J., S. van de Par, and A. Kohlrausch. 2001b. Binaural processing model based on contralateral inhibition. II. Dependence on spectral parameters. *Journal of the Acoustical Society of America* 110: 1089–1103.
- Breebaart, J., S. van de Par, and A. Kohlrausch. 2001c. Binaural processing model based on contralateral inhibition. III. Dependence on temporal parameters. *Journal of the Acoustical Society of America* 110: 1117–1125.
- Bregman, A.S. 1990. *Auditory Scene Analysis*. Cambridge, MA: MIT Press.
- Brown, G.J., and M.P. Cooke. 1994. Computational auditory scene analysis. *Computer Speech and Language* 8: 297–336.
- Brown, G.J., S. Harding, and J.P. Barker, 2006. Speech separation based on the statistics of binaural auditory features. In *Proceedings of IEEE International Conference Acoustical, Speech, and Signal Processing*, vol. V, 949 – 952.
- Brown, G.J., and K.J. Palomäki. 2011. A computational model of binaural speech recognition: Role of across-frequency vs. within-frequency processing and internal noise. *Speech Communication* 53: 924–940.
- Burkhard, M.D., and R.M. Sachs. 1975. Anthropometric manikin for acoustic research. *Journal of the Acoustical Society of America* 58: 214–222.
- Cantu, M. 2018. Sound source segregation of multiple concurrent talkers via short-time target cancellation. Ph.D. thesis, Boston University.
- Cho, B.J., H. Kwon, J.-W. Cho, C. Kim, R.M. Stern, and H.-M. Park. 2016. A subband-based stationary-component suppression method using harmonics and power ratio for reverberant speech recognition. *IEEE Signal Processing Letters* 23 (6): 780–784.

- Colburn, H.S. 1969. Some physiological limitations on binaural performance. Ph.D. thesis, Massachusetts Institute of Technology.
- Colburn, H.S. 1973. Theory of binaural interaction based on auditory-nerve data. I. general strategy and preliminary results on interaural discrimination. *Journal of the Acoustical Society of America* 54: 1458–1470.
- Colburn, H.S., and N.I. Durlach. 1978. Models of binaural interaction. In *Hearing*, ed. E.C. Carterette, and M. P. Friedmann, Vol. IV of Handbook of Perception, Chap. 11, 467–518. New York: Academic Press
- Colburn, H.S., and A. Kulkarni. 2005. Models of sound localization. In *Sound Source Localization*, eds. R. Fay, and T. Popper, *Springer Handbook of Auditory Research*, Chap. 8, 272–316. Springer-Verlag
- Cooke, M., P. Green, L. Josifovski, and A. Vizinho. 2001. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication* 34: 267–285.
- Cooke, M.P., and D. P.W. Ellis. 2001. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication* 35, 141–177.
- Davis, S.B., and P. Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28: 357–366.
- Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* 39: 1–38.
- DeSimio, M.P., T.R. Anderson, and J.J. Westerkamp. 1996. Phoneme recognition with a model of binaural hearing. *IEEE Transactions on Speech and Audio Processing* 4: 157–166.
- Dietz, M., J.H. Lestang, P. Majdak, R.M. Stern, T. Marquardt, S.D. Ewert, W.M. Hartmann, and D.F.M. Goodman. 2017. A framework for testing and comparing binaural models. *Hearing Research* 360: 92–106.
- Dietz, M., T. Marquardt, N.H. Salminen, and D. McAlpine. 2013. Emphasis of spatial cues in the temporal fine structure during the rising segments of amplitude-modulated sounds. *Proceedings of the National Academy of Sciences of the United States of America* 110: 15151–15156.
- Domnitz, R.H., and H.S. Colburn. 1976. Analysis of binaural detection models for dependence on interaural target parameters. *Journal of the Acoustical Society of America* 59: 599–601.
- Domnitz, R.H., and H.S. Colburn. 1977. Lateral position and interaural discrimination. *Journal of the Acoustical Society of America* 61: 1586–1598.
- Droppo, J. 2013. Feature compensation. In *Techniques for Noise Robustness in Automatic Speech Recognition*, ed. T. Virtanen, B. Raj, and R. Singh, Chap. 9. Wiley
- Durlach, N.I. 1963. Equalization and cancellation theory of binaural masking level differences. *Journal of the Acoustical Society of America* 35 (8): 1206–1218.
- Durlach, N.I. 1972. Binaural signal detection: Equalization and cancellation theory. In *Foundations of Modern Auditory Theory*, vol. 2, ed. J.V. Tobias, 369–462. New York: Academic Press.
- Durlach, N.I., and H.S. Colburn. 1978. Binaural phenomena. In *Hearing*, ed. E.C. Carterette, and M.P. Friedman, 365–466., Vol. IV of Handbook of Perception New York: Academic Press.
- Faller, C., and J. Merimaa. 2004. Sound localization in complex listening situations: Selection of binaural cues based on interaural coherence. *Journal of the Acoustical Society of America* 116 (5): 3075–3089.
- Fan, N., J. Du, and L.-R. Dai. 2016. A regression approach to binaural speech segregation via deep neural networks. In *Proceedings of IEEE International Symposium on Chinese Spoken Language Processing*, 116–120.
- Flanagan, J.L., J.D. Johnston, R. Zahn, and G.W. Elko. 1985. Computer-steered microphone arrays for sound transduction in large rooms. *Journal of the Acoustical Society of America* 78: 1508–1518.
- Gaik, W. 1993. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *Journal of the Acoustical Society of America* 94: 98–110.
- Gardner, B., and K. Martin. 1994. HRTF measurements of a KEMAR dummy-head microphone. Technical Report 280. Available online at <http://sound.media.mit.edu/KEMAR.html>.

- Gilkey, R.H., and Anderson, T.A. (eds.). 1997. *Binaural and Spatial Hearing in Real and Virtual Environments*. Psychology Press.
- Gold, B., N. Morgan, and D. Ellis. 2011. *Speech and Audio Signal Processing*, 2nd ed. Wiley Interscience.
- Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press.
- Harding, S., J. Barker, and G.J. Brown. 2006. Mask estimation for missing data speech recognition based on statistics of binaural interaction. *IEEE Transactions on Speech and Audio Processing* 14: 58–67.
- Hartung, K., and C. Trahiotis. 2001. Peripheral auditory processing and investigations of the “precedence effect” which utilize successive transient stimuli. *Journal of the Acoustical Society of America* 110 (3): 1505–1513.
- Hawley, M.L., R.Y. Litovsky, and H.S. Colburn. 1999. Speech intelligibility and localization in a multi-source environment. *Journal of the Acoustical Society of America* 105: 3436–3448.
- Haykin, S. 2018. *Neural Networks And Learning Machines*, 3rd ed. Springer.
- Hermansky, H. 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America* 87 (4): 1738–1752.
- Hermansky, H., D.P.W. Ellis, and S. Sharma. 2000. Tandem connectionist feature extraction for conventional hmm systems. In *Proceedings of the IEEE ICASSP*, 1635–1638.
- Hermansky, H., and N. Morgan. 1994. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing* 2: 578–589.
- Hinton, G., L. Deng, D. Yu, G.E. Dahl, and Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* 29, 82–97.
- Jeffress, L.A. 1948. A place theory of sound localization. *Journal of Comparative Physiology, Psychology* 41: 35–39.
- Jeub, M., M. Dorbecker, and P. Vary. 2011a. Semi-analytical model for the binaural coherence of noise fields. *IEEE Signal Processing Letters* 18 (3): 197–200.
- Jeub, M., C. Nelke, C. Beaugeant, and P. Vary. 2011b. Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals. In *Proceedings of the 19th European Signal Processing Conference*.
- Jeub, M., M. Schafer, T. Esch, and P. Vary. 2010. Model-based dereverberation preserving binaural cues. *IEEE Transactions on Audio, Speech, and Language Processing* 18 (7): 1732–1745.
- Jeub, M., M. Schafer, and P. Vary. 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proceedings on 16th International Conference on Digital Signal Processing*, 1–5.
- Jiang, Y., D. Wang, R. Liu, and Z. Feng. 2014. Binaural classification for reverberant speech segregation using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22 (12): 2112–2121.
- Johnson, D.H., and D.E. Dudgeon. 1993. *Array Signal Processing: Concepts and Techniques*. Englewood Cliffs NJ: Prentice-Hall.
- Kates, J.M. 1991. A time-domain digital cochlear model. *IEEE Transaction on Signal Processing* 39: 2573–2592.
- Kim, C., C. Khawand, and R.M. Stern. 2012. Two-microphone source separation algorithm based on statistical modeling of angle distributions. In *Proceedings of the IEEE International Conference Acoustical, Speech and Signal Processing*.
- Kim, C., K. Kumar, B. Raj, and R.M. Stern. 2009. Signal separation for robust speech recognition based on phase difference information obtained in the frequency domain. In *Proceedings of the Interspeech Conference*.
- Kim, C., K. Kumar, and R.M. Stern. 2011. Binaural sound source separation motivated by auditory processing. In *Proceedings of the Interspeech Conference*, Prague, Czech Republic, vol. 23, 780–784.

- Kim, C., and R.M. Stern. 2010. Nonlinear enhancement of onset for robust speech recognition. In *Proceedings of the Interspeech Conference*. Makuhari, Japan
- Kim, C., and R.M. Stern. 2016. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 24(7), 1315–1329.
- Kim, C., R.M. Stern, K. Eom, and J. Kee. 2010. Automatic selection of thresholds for signal separation algorithms based on interaural delay. In *Proceedings of the Interspeech Conference*. Makuhari, Japan.
- Kohonen, T. 1989. The neural phonetic typewriter. *IEEE Computer Magazine*, 11–22.
- Kohlrausch, A., J. Braasch, D. Kolossa, and J. Blauert. 2013. An introduction to binaural processing. In *The Technology of Binaural Listening*, ed. J. Blauert., Springer and ASA Press.
- Kumatani, K., J. McDonough, and B. Raj. 2012. Microphone array processing for robust speech recognition. *IEEE Signal Processing Magazine* 29 (6): 127–140.
- Lindemann, W. 1986a. Extension of a binaural cross-correlation model by contralateral inhibition. I. simulation of lateralization for stationary signals. *Journal of the Acoustical Society of America* 80: 1608–1622.
- Lindemann, W. 1986b. Extension of a binaural cross-correlation model by contralateral inhibition. II. the law of the first wavefront. *Journal of the Acoustical Society of America* 80: 1623–1630.
- Lippmann, R.P. 1987. An introduction to computing with neural nets. *IEEE ASSP Magazine* 4 (2): 4–22.
- Lippmann, R.P. 1989. Review of neural networks for speech recognition. *Neural Computation* 1 (1): 1–38.
- Litovsky, R.Y., S.H. Colburn, W.A. Yost, and S.J. Guzman. 1999. The precedence effect. *Journal of the Acoustical Society of America* 106: 1633–1654.
- Lyon, R.F. 1984. Computational models of neural auditory processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing of the International Conference on Acoustics, Speech and Signal Processing*, 36.1.1–36.1.4.
- Mandel, M.I., R.J. Weiss, and D.P.W. Ellis. 2010. Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing* 18 (2): 382–394.
- Martin, K.D. 1997. Echo suppression in a computational model of the precedence effect. In *Proceedings of the IEEE Mohonk Workshop on Applications of Signal Processing to Acoustics and Audio*.
- May, T., S.V.D. Par, and A. Kohlrausch. 2012. A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation. *IEEE Transactions on Audio, Speech, and Language Processing* 20: 108–121.
- May, T., S. van de Par, and A. Kohlrausch. 2011. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech, and Language Processing* 19 (1): 1–13.
- McGovern, S.G. 2004. Room impulse response generator (MATLAB code). <http://www.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator>.
- Mehrgardt, S., and V. Mellert. 1977. Transformation characteristics of the external human ear. *Journal of the Acoustical Society of America* 61: 1567–1576.
- Menon, A. 2018. Robust recognition of binaural speech signals using techniques based on human auditory processing. Ph.D. thesis, Carnegie Mellon University.
- Mi, J., and H.S. Colburn. 2016. A binaural grouping model for predicting speech intelligibility in multitalker environments. *Trends in Hearing* 20: 1–12.
- Mi, J., M. Groll, and H.S. Colburn. 2017. Comparison of a target-equalization-cancellation approach and a localization approach to source separation. *Journal of the Acoustical Society of America* 142 (5): 2933–2941.
- Miao, Y., and F. Metze. 2017. End-to-end architectures for speech recognition. In *New Era for Robust Speech Recognition: Exploiting Deep Learning*, ed. Watanabe, S., M. Delcroix, F. Metze, and J.R. Hershey, 299–323. Springer International Publishing

- Mitra, V., H. Franco, R. Stern, J.V. Hout, L. Ferrer, M. Graciarena, W. Wang, D. Vergyri, A. Alwan, and J.H.L. Nansen. 2017. Robust features in deep learning-based speech recognition. In *New Era for Robust Speech Recognition: Exploiting Deep Learning*, ed. Watanabe, S., M. Delcroix, F. Metze, and J.R. Hershey, 183–212. Springer International Publishing
- Moore, B.C.J. 2012. *An Introduction to the Psychology of Hearing*, 6th ed. Bingley UK, London: Emerald Group Publishing Ltd.
- Moreno, P.J., B. Raj, and R.M. Stern. 1996. A vector Taylor series approach for environment-independent speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 733–736
- Nielsen, M. 2016. *Neural Networks and Deep Learning*. <http://neuralnetworksanddeeplearning.com/>.
- Osman, E. 1971. A correlation model of binaural masking level differences. *Journal of the Acoustical Society of America* 50: 1494–1511.
- Palomäki, K.J., G.J. Brown, and D.L. Wang. 2004. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Communication* 43 (4): 361–378.
- Park, H.-M., and R.M. Stern. 2009. Spatial separation of speech signals using continuously-variable weighting factors estimated from comparisons of zero crossings. *Speech Communication Journal* 51 (1): 15–25.
- Patterson, R.D., I. Nimmo-Smith, J. Holdsworth, and P. Rice. 1988. *An efficient auditory filterbank based on the gammatone function*, Applied Psychology Unit (APU) Report 2341. Cambridge UK
- Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77 (2): 257–286.
- Rabiner, L.R., and B.-H. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall.
- Raj, B., M.L. Seltzer, and R.M. Stern. 2004. Reconstruction of missing features for robust speech recognition. *Speech Communication* 43 (4): 275–296.
- Raj, B., and R.M. Stern. 2005. Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine* 22 (5): 101–115.
- Rickard, S. 2007. The DUET blind source separation algorithm. In *Blind Speech Separation*, ed. Makino, S., T. Lee, and H.E. Sawada. New York: Springer-Verlag.
- Roman, N., S. Srinivasan, and D. Wang. 2006. Binaural segregation in multisource. *Journal of the Acoustical Society of America* 120: 4040–4051.
- Roman, N., D.L. Wang, and G.J. Brown. 2003. Speech segregation based on sound localization. *Journal of the Acoustical Society of America* 114 (4): 2236–2252.
- Rosenblatt, R. 1959. *Principles of Neurodynamics*. New York: Spartan Books.
- Schroeder, M.R. 1977. New viewpoints in binaural interactions. In *Psychophysics and Physiology of Hearing*, ed. Evans, E.F. and J.P. Wilson, 455–467. London: Academic Press
- Shamma, S.A., N. Shen, and P. Gopalswamy. 1989. Binaural processing without neural delays. *Journal of the Acoustical Society of America* 86: 987–1006.
- Shao, Y., and D.L. Wang. 2008. Robust speaker identification using auditory features and computational auditory scene analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1589–1592
- Srinivasan, S., M. Roman, and D. Wang. 2006. Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication* 48: 1486–1501.
- Stecker, G.C., J.D. Ostreicher, and A.D. Brown. 2013. Temporal weighting functions for interaural time and level differences. III. Temporal weighting for lateral position judgments. *Journal of the Acoustical Society of America* 134: 1242–1252.
- Stern, R.M., and H.S. Colburn. 1978. Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position. *Journal of the Acoustical Society of America* 64: 127–140.
- Stern, R.M., and Trahiotis, C. 1995. Models of binaural interaction. In *Hearing*, ed. Moore, B.C.J., Handbook of Perception and Cognition, 2 ed, Chap. 10, 347–386. New York: Academic.

- Stern, R.M., and C. Trahiotis. 1996. Models of binaural perception. In *Binaural and Spatial Hearing in Real and Virtual Environments*, ed. Gilkey, R. and T.R. Anderson, Chap. 24, 499–531. Lawrence Erlbaum Associates
- Stern, R.M., D. Wang, and G.J. Brown. 2006. Binaural sound localization. In *Computational Auditory Scene Analysis*, ed. Wang, D., and G.J. Brown, Chap. 5. Wiley-IEEE Press
- Stern, R.M., A.S. Zeiberg, and C. Trahiotis. 1988. Lateralization of complex binaural stimuli: a weighted image model. *Journal of the Acoustical Society of America* 84: 156–165.
- Stevens, S.S., J. Volkman, and E. Newman. 1937. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America* 8 (3): 185–190.
- Stockham, T.G., T.M. Cannon, and R.B. Ingrebretsen. 1975. Blind deconvolution through digital signal processing. *Proceedings of the IEEE* 63 (4): 678–692.
- Thiergart, O., G. Del Galdo, and E.A. Habets. 2012. Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 309–312.
- Trahiotis, C., L.R. Bernstein, R.M. Stern, and T.N. Buell. 2005. Interaural correlation as the basis of a working model of binaural processing: An introduction. In *Sound Source Localization*, ed. R. Fay, and T. Popper, 238–271., *Springer Handbook of Auditory Research*. Heidelberg: Springer-Verlag.
- Van Trees, H.L. 2004. *Detection, Estimation, and Modulation Theory: Optimum Array Processing*. Wiley.
- Virtanen, T., B. Raj, and R. Singh, eds. 2012. *Noise-Robust Techniques for Automatic Speech Recognition*. Wiley.
- Wallach, H.W., E.B. Newman, and M.R. Rosenzweig. 1949. The precedence effect in sound localization. *American Journal of Psychology* 62: 315–337.
- Wan, R., N.I. Durlach, and H.S. Colburn. 2010. Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. *Journal of the Acoustical Society of America* 128: 3678–3690.
- Wan, R., N.I. Durlach, and H.S. Colburn. 2014. Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers. *Journal of the Acoustical Society of America* 136: 768–776.
- Wang, D., and G.J. Brown, eds. 2006. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.
- Wang, D.L., and J. Chen. 2018. Supervised speech separation based on deep learning: An overview. *IEEE Transactions on Audio, Speech, and Language Processing* 26: 1702–1726.
- Wang, Y., and D.L. Wang. 2013. Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing* 21: 1381–1390.
- Watanabe, S., M. Delcroix, F. Metze, and J.R. Hershey, eds. 2017. *New Era for Robust Speech Recognition: Exploiting Deep Learning*. Springer International.
- Westermann, A., J.M. Buchholz, and T. Dau. 2013. Binaural dereverberation based on interaural coherence histograms. *The Journal of the Acoustical Society of America* 133 (5): 2767–2777.
- Wightman, F.L., and D.J. Kistler. 1989a. Headphone simulation of free-field listening. I: Stimulus synthesis. *The Journal of the Acoustical Society of America* 85: 858–867.
- Wightman, F.L., and D.J. Kistler. 1989b. Headphone simulation of free-field listening. II: Psychophysical validation. *Journal of the Acoustical Society of America* 87: 868–878.
- Wightman, F.L., and D.J. Kistler. 1999. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America* 105 (5): 2841–2853.
- Woodruff, J., and D.L. Wang. 2013. Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues. *IEEE Transactions on Audio, Speech, and Language Processing* 21: 806–815.
- Yost, W.A. 1981. Lateral position of sinusoids presented with intensive and temporal differences. *Journal of the Acoustical Society of America* 70: 397–409.
- Yost, W.A. 2013. *Fundamentals of Hearing: An Introduction*, 5th ed. Burlington MA: Academic Press.

- Yu, Y., W. Wang, and P. Han. 2016. Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks. *EURASIP Journal on Audio, Speech, and Music Processing* 2016: 1–18.
- Zhang, X., M.G. Heinz, I.C. Bruce, and L.H. Carney. 2001. A phenomenological model for the response of auditory-nerve fibers: I. nonlinear tuning with compression and suppression. *Journal of the Acoustical Society of America* 109: 648–670.
- Zhang, X., and D. Wang. 2017. Deep learning based binaural speech separation in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (5): 1075–1084.
- Zheng, C., A. Schwarz, W. Kellermann, and X. Li. 2015. Binaural coherent-to-diffuse-ratio estimation for dereverberation using an ITD model. In *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO)*, 1048–1052.
- Zilany, M.S.A., I.C. Bruce, P.C. Nelson, and L.H. Carney. 2009. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. *Journal of the Acoustical Society of America* 125: 2390–2412.
- Zurek, P.M. 1993. Binaural advantages and directional effects in speech intelligibility. In *Acoustical Factors Affecting Hearing Aid Performance*, ed. G.A. Studebaker, and I. Hochberg. Boston: Allyn and Bacon.
- Zurek, P.M., R.L. Freyman, and U. Balakrishnan. 2004. Auditory target detection in reverberation. *Journal of the Acoustical Society of America* 115 (4): 1609–1620.

Modeling Binaural Speech Understanding in Complex Situations



Mathieu Lavandier and Virginia Best

Abstract This chapter reviews binaural models available to predict speech intelligibility for different kinds of interference and in the presence of reverberation. A particular effort is made to quantify their performances and to highlight the a priori knowledge they require in order to make a prediction. In addition, cognitive factors that are not included in current models are considered. The lack of these factors may limit the ability of current models to predict speech understanding in real-world listening situations fully.

1 Introduction

In order to understand speech in a noisy environment or when many people are talking at the same time, a listener must hear out the target speech from the background noise and segregate the target talker from competing talkers. This challenging task becomes even more complicated in built environments, where the auditory system also has to cope with the effects of reverberation.

The following section reviews some behavioral data that describe the masking of speech in listeners with normal hearing, cues that can alleviate masking, and some relevant effects of reverberation. Then an overview of several models that have been proposed to predict speech understanding in complex binaural listening situations

M. Lavandier (✉)

Univ Lyon, ENTPE, Laboratoire Génie Civil Et Bâtiment, Rue M. Audin, 69518 Vaulx-en-Velin
Cedex, France

e-mail: mathieu.lavandier@entpe.fr

V. Best

Department of Speech, Language and Hearing Sciences, Boston University, 635 Commonwealth
Ave, Boston, MA 02215, USA

e-mail: ginbest@bu.edu

© Springer Nature Switzerland AG 2020

J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,

https://doi.org/10.1007/978-3-030-00386-9_19

is given. One of the main factors determining speech intelligibility in noise is the Signal-to-Noise Ratio (SNR) at which sounds are presented to the listener. The binaural models presented below fall under the broad categories of SNR-based models, modulation-based models, correlation-based models, and segregation-based models. Finally, some issues with the practical application of these models are discussed, and several cognitive factors are considered that are currently unaccounted for.

2 Overview of Relevant Behavioral Data

2.1 *Masking of Speech*

Several distinct kinds of interference can degrade speech intelligibility in situations containing competing sounds. Energetic Masking (EM) describes interference that occurs when the competing sounds overlap¹ acoustically with the target and render it inaudible. EM is typically studied using broadband maskers that exert their masking effect relatively uniformly across time and frequency. Many acoustical cues can reduce the EM caused by a competing sound, as described in a recent review chapter by Culling and Stone (2017). For example, modulations in the temporal envelope of a noise masker can reduce its effectiveness by providing improvements in SNR in moments when the masker envelope is low (Bronkhorst and Plomp 1992; Festen and Plomp 1990). This ability to exploit temporal fluctuations in the level of the masker is called dip listening, listening in the gaps, or glimpsing (Cooke 2006). In addition, differences in the source location or harmonic structure of competing sounds can reduce EM.

Informational Masking (IM) describes interference that cannot be explained in energetic terms and is thought to occur more centrally (see the recent review by Kidd and Colburn 2017). In the context of speech intelligibility, it describes the additional interference that occurs when similar competing talkers mask a speech target. IM can refer to both difficulties in segregating speech mixtures (i.e., determining which parts belong to the target) and difficulties in attending to the desired source in the mixture (i.e., overcoming confusion or distraction).

¹The term “overlap” is used loosely here, because both forward masking and upward spread of masking are still considered a form of EM. Forward masking refers to the masking of a sound by a preceding sound. In this case, there is no temporal overlap between the target and the competing sound that can precede the target by up to 200 ms. Upward spread of masking refers to the masking of a high-frequency sound by a low-frequency sound. In this case, there is no spectral overlap between target and masker.

2.2 *Spatial Release from Masking*

Possessing two ears is useful for understanding speech in noise: a competing sound source causes less masking when it is separated spatially from the target speech (Plomp 1976; Hawley et al. 2004). Spatial Release from Masking (SRM) is often measured by subtracting Speech Reception Thresholds (SRTs) measured in a co-located condition from SRTs measured in a separated condition.

For situations dominated by EM, SRM is thought to be based on two main mechanisms (Bronkhorst and Plomp 1988): better-ear listening and binaural unmasking. Better-ear listening results from differences in the power level of the sound at the two ears (Interaural Level Differences, ILDs). For sources located to one side of a listener, the sound level is reduced at the far ear—the ear for which the head throws an acoustic shadow—creating an ILD. Target and interferers at different locations often produce different ILDs, so one ear will usually offer a better SNR than the other, and listeners can simply use the information coming from whichever ear offers the better SNR. Binaural unmasking relies on differences in the timing of the sound at the two ears (Interaural Time Differences, ITDs). For lateral sources the sound arrives later at the contralateral ear, because the sound must travel farther from the source to this ear, thus generating an ITD between both ears. Differences in the ITDs generated by the target and an interferer facilitate binaural unmasking, a condition in which the central auditory nervous system can “cancel” sounds generated by the interferer to some extent (Equalization-Cancellation (E-C) theory; Durlach 1972). This way the E-C mechanism can improve the internal SNR. Binaural unmasking is also sometimes called binaural interaction or binaural squelch. Licklider (1948) showed that speech intelligibility in noise diminishes when the interaural coherence of the noise is reduced. This effect can be explained by the E-C theory, which predicts that a less correlated masker cannot be fully equalized at the two ears, and hence cannot be fully canceled, resulting in more masking and lower speech intelligibility.

Recent work suggests that in complex listening situations where the interference comes from multiple fluctuating interferers, a more sophisticated definition of the better-ear advantage might be required. For example, Schoenmaker et al. (2017) suggested that the better ear can be defined based on the number of good target “glimpses” available at each ear, and the better-ear advantage can be estimated based on pooled local SNRs (i.e., calculated within relatively narrow frequency channels and short time windows) rather than the global SNR. In a related study, Brungart and Iyer (2012) suggested that the better ear could even be chosen on a moment-by-moment basis, independently in each frequency channel. They showed that the benefit estimated from such a mechanism could predict performance in binaural mixtures with two symmetrically placed maskers. Later work, however, showed that this approach underestimates the SRM that is observed for mixtures that include many more maskers (Lingner et al. 2016) or maskers that cause a lot of IM (Glyde et al. 2013a). It is not established yet whether there is a “true” binaural mechanism that involves switching across ears to get the most relevant information (best SNRs), or whether two monaural mechanisms provide the SNRs simultaneously at both

ears. Culling and Mansell (2013) showed that the use of a better ear that alternated between the ears was highly dependent on the switching rate, suggesting a rather sluggish mechanism.

Differences in spatial location between a target and interferers can largely reduce the effect of IM, but likely via different mechanisms than those driving spatial release from EM. Spatial separation reduces the uncertainty about how to disentangle the different sources in a mixture and provides a substrate for selective attention. Perhaps the best illustration of spatial release from IM comes from studies that have found large amounts of SRM for stimuli in which there is no change in the EM. For example, Arbogast et al. (2002) presented target and masker speech that were restricted to mutually exclusive narrow frequency bands, thus minimizing the spectral overlap. For these stimuli, an SRM of around 13 dB was observed. In a similar paradigm, Best et al. (2011) presented target and masker sentences that were temporally interleaved on a word-by-word basis, so that no simultaneous masking could occur. For these stimuli, a robust SRM of around 30% points was observed. Freyman et al. (1999, 2001) used the precedence effect to create the illusion of spatial separation between competing talkers and found a large SRM despite a small *increase* in EM. For natural speech mixtures, the primary benefit of spatial separation could come from grouping based on binaural cues, rather than binaural unmasking as described above for speech in noise (Schoenmaker et al. 2016). Consistent with this notion, it appears that any cue that provides the perception of spatial separation is sufficient to provide an SRM for speech mixtures: ITDs and ILDs alone (Best et al. 2013; Glyde et al. 2013b), and even monaural spectral cues associated with separation in distance and elevation (Martin et al. 2012; Brungart and Simpson 2002).

2.3 *Influence of Non-spatial Cues*

In addition to spatial separation, there are several other (non-spatial) cues that can provide a release from masking of speech. For example, differences in voice characteristics between the competing talkers can greatly reduce masking and improve target intelligibility. While voices can differ along many dimensions, much experimental work has focused on differences in fundamental frequency (F0). Brokx and Nooteboom (1982) used monotonized speech to demonstrate that intelligibility of a target (F0 = 100 Hz) systematically improved as the F0 of the masker was increased from 100 to 200 Hz.

The energetic component of this effect has been explored in experiments using harmonic complex maskers. Speech intelligibility is improved when the difference in F0 between the speech target and a harmonic masker is increased (Deroche and Culling 2011). This benefit can be influenced greatly by F0 fluctuations across time, which are found in intonated speech. Deroche and Culling (2011) observed a detrimental effect of sinusoidally modulating the F0 of a harmonic complex masker; in contrast, the same F0 modulation applied to the target voice had no impact on its intelligibility. For natural F0 fluctuations like those found in intonated speech,

Leclère et al. (2017) showed that, when a harmonic complex masker was intonated, the benefit of a difference in mean F0 between target and masker was abolished. When only the target was intonated, this benefit was also largely reduced. However, this probably had a very different cause: in addition to providing prosodic cues that facilitate intelligibility regardless of masking (Binns and Culling 2007), natural F0 fluctuations in the target provided instantaneous F0 differences which cause a release from masking that subsequently reduces the benefit of a difference in mean F0 (the F0-segregation mechanism being potentially at ceiling). This suggests that for situations involving primarily EM, there is a differential role for natural F0 fluctuations: those of the target voice are beneficial whereas those of the masker are detrimental.

Several mechanisms have been proposed for the release from masking afforded by differences in F0 (see review by Culling and Stone 2017). One mechanism involves “spectral glimpsing” or the accessing of target information in between the resolved harmonics of the masker (Deroche et al. 2014). Another mechanism invokes the harmonic cancellation theory (de Cheveigné et al. 1995): The auditory system may identify the harmonic structure of the masker in order to cancel it and improve the SNR when the target and masker differ in F0.

Although difficult to isolate from EM effects, there are several indications that differences in voice between competing talkers can also provide a release from IM. For example, voice differences have been examined using tasks designed to maximize IM, in which listeners must attend to one of two or more highly similar and synchronized sentences. Substantial improvements in performance are observed when the masker talkers are different talkers than the target talker, especially if they differ in sex from the target (e.g., Brungart 2001; Brungart et al. 2001). Further evidence for non-energetic influences of voice cues comes from studies that have reported particular benefits of familiar talkers in speech-on-speech tasks (e.g., Johnsrude et al. 2013; Souza et al. 2013). Finally, it is worth noting that release from IM afforded by voice cues (and other non-spatial cues) can interact with the release afforded by spatial cues. For instance, there are examples in the literature showing that spatial separation provides a much-reduced benefit for different-sex or time-reversed maskers in which IM is already greatly reduced (e.g., Best et al. 2013; Xia et al. 2015).

2.4 *Effects of Reverberation*

When communicating in noisy rooms, reverberation has several effects on speech intelligibility. First, reverberation exerts a well-known temporal smearing effect on the target speech, which occurs even in quiet. Having two ears may ameliorate this smearing effect on target intelligibility. This binaural de-reverberation has been shown to slightly improve intelligibility for reverberant speech in quiet (Moncur and Dirks 1967; Nábělek and Robinson 1982) and in the presence of a noise interferer (Lavandier and Culling 2008).

Sound reflections in rooms can also reduce the possibility for dip listening (Bronkhorst and Plomp 1990; George et al. 2008; Beutelmann et al. 2010; Collin and

Lavandier 2013). They will tend to reduce the envelope modulations of the masker, thus increasing the masking by filling in the gaps through which the target could be heard. Another effect of reverberation that influences speech intelligibility in noise is the modification of source spectra at the ears of the listener by room “coloration”, which results from both the constructive/destructive interferences of sound reflections and the frequency-dependent absorption characteristics of room materials. The spectrum produced by each source at each ear depends on the ear and source positions within the room. Thus, coloration influences intelligibility by determining the frequency-dependent SNR at the ears.

Several studies have shown that reverberation reduces SRM (Plomp 1976; Culling et al. 2003; Beutelmann and Brand 2006). Sound reflections traveling around the listener reduce ILDs, thus critically impairing better-ear listening (Plomp 1976). Moreover, because these reflections are typically not identical at the two ears, they impair binaural unmasking by decorrelating the interfering sound at the two ears (Lavandier and Culling 2008). Room reflections also modify the signal ITDs, further affecting binaural unmasking, which depends on the ITD differences between target and interferer (Lavandier and Culling 2010). Although there are only limited data on the issue, spatial release from IM appears to be more robust to reverberation than is spatial release from EM (Kidd et al. 2005b). This probably reflects the fact that it is the *perception* of spatial separation, rather than a specific acoustical cue, that drives release from IM.

Reverberation can also be detrimental to the segregation of voices based on F0 (Culling et al. 1994, 2003; Deroche and Culling 2011). When a harmonic masker has a fluctuating F0, then the delayed versions of the masker associated with room reflections have a different F0 than the masker version carried by the direct sound. As a result, reverberation makes the masker less harmonic and fills in the spectral gaps in between its resolved frequency components. Room reflections should impair both spectral glimpsing and harmonic cancellation.

3 Review of Binaural Intelligibility Models

Several models have been proposed to predict speech understanding in binaural listening situations. They are presented below along with, whenever available, a quantitative evaluation of their performances. This evaluation generally relies on the correlation between measured and predicted SRTs (*corr*), the mean absolute prediction error (absolute differences between measured and predicted SRTs averaged across conditions/data points, *mean err*), the root-mean-square prediction error (*rms err*) or the maximum absolute prediction error (*max err*).

The models are gathered here into broad categories, based on the information they are considering in the signals (e.g., energy, modulations, glimpses) that is assumed to be the key information regarding intelligibility, and based on their required inputs. For example, the SNR-based models will require the target/signal (S) and interferer/noise (N) available separately at the ears in order to compute SNRs. The different model

categories were developed rather independently from each other, as researchers tried to evaluate the potential of each method/type of information. The different versions of a model within each category were generally proposed to extend the scope of the original model (e.g., the situations or types of interferer it can deal with), or to resolve limitations of the previous model versions.

3.1 SNR-Based Models

The Speech Intelligibility Index (SII; ANSI S3.5 1997)—a successor of the Articulation Index (AI; Kryter 1962)—is a widely used (monaural) indicator to predict intelligibility in noise. It is calculated by computing the SNRs in different frequency bands covering the speech spectrum (below 10 kHz), applying a simple weighting to these ratios to give more importance to the frequency bands that are the most important for understanding speech (400–4400 Hz), and summing the weighted ratios across frequency. The SII can also take into account upward spread of masking (see Footnote 1) and hearing threshold. It predicts the proportion of the total speech information that is audible and available to the listener on a range between 0 and 1 (Rhebergen and Versfeld 2005). When considering intelligibility in the presence of non-stationary noises, dip listening needs to be taken into account. To predict the monaural advantage of dip listening, Rhebergen and Versfeld (2005) proposed a short-time version of the SII. This extended SII decomposes the signals using short-time frames, computes the SII within each frame, and averages the SII predictions across frames. In order to consider the pauses/envelope modulations in the target speech as relevant information for its intelligibility, the model considers the average level of the target across time rather than its instantaneous level within short-time frames. Computing the SNR with the actual speech waveform would mistakenly lead to a reduced effective SNR within the target pauses; thus implicitly considering these pauses as an absence of information. Peaks in the interferer signal induce an increase of masking, whereas pauses induce a decrease of masking. Therefore, the model needs to consider interfering energy as a function of time, and target speech energy averaged across time—see also Collin and Lavandier (2013).

The approach of Rhebergen and Versfeld predicted monaural SRTs measured with stationary noise, speech-modulated noises, and interrupted noise reasonably well. The model's agreement with data was taken to postulate that “*average speech intelligibility in fluctuating noise can be modeled by averaging the amount of speech information across time*” (Rhebergen and Versfeld 2005). The evaluation of model performance in terms of prediction errors between measured and predicted SRTs was not always explicit. For speech-modulated noises, the maximum prediction error *max err* seemed to be at least 2.3 dB, while prediction performance was worse for sinusoidally-modulated noises (*max err* of 3.9 dB). This monaural model was later refined to take into account forward masking (Rhebergen et al. 2006). This addition proved useful to better predict the effect of dip listening for noise maskers with large silent gaps and abrupt offsets, but it did not improve prediction for sinusoidally-

modulated noises without silent periods, and its relevance for maskers containing speech modulations is not completely clear. Overall, across nineteen noise conditions including various interrupted, saw-tooth and sinusoidally-modulated noises, *max err* was 2.9 dB for the extension with forward masking (excluding a condition with interrupted noise modulated at 4 Hz, which increased *max err* to 12 dB), while *max err* was up to 10.7 dB for the non-extended version of the model.

Computational SNR Models

One kind of binaural SNR-based model, referred to here as “computational”, uses a direct implementation of an E-C process as described by Durlach (1972). Beutelmann and Brand (2006) developed a model predicting the intelligibility of a speech target in the presence of a stationary-noise interferer in rooms (Fig. 1). Simulated stimuli at the ears are first processed through a gammatone filterbank and an E-C stage, then re-synthesized with the binaurally-enhanced SNR, and then the SII is computed to evaluate intelligibility. For each frequency band of the filterbank, an E-C mechanism is implemented: the left- and right-ear signals are attenuated and delayed with respect to each other (equalization), and then cancellation is simulated by subtracting the right channel from the left channel. Different delays and attenuations are tested in the equalization step, and those maximizing the effective SNR after cancellation are selected. Cancellation is then applied to the signals before re-synthesis. If the effective SNR after E-C processing is lower than the SNR at the left or right ear, the signals at the ear having the best SNR are selected for re-synthesis and SII calculation. Beutelmann and Brand (2006) obtained very good agreement between model predictions and listening test data involving single noise interferers in three different rooms (overall *corr* of 0.95, *mean err* of 1.6 dB in anechoic conditions, 0.5 dB in an office, 0.3 dB in a cafeteria).

Beutelmann et al. (2010) applied the method of Rhebergen and Versfeld (2005) to extend their binaural model to predict SRM for a non-stationary noise. The model is then used within short-time frames before averaging the resulting predictions. The latter were compared to SRTs measured with stationary, 1-voice modulated and 20-voice modulated noises, simulated in four virtual rooms (one being anechoic) and three configurations (varying the source distance to the listener and the azimuthal separation of sources). Depending on the noise type, *corr* values ranging from 0.80 to 0.93 were obtained, with an overall *mean err* of 3 dB. In another recent study (Ewert et al. 2017), results from this model were compared to empirical data collected using a range of masking sounds that varied in their spectro-temporal properties and the amount of IM they were expected to produce. The binaural cues available to the listeners were also manipulated during the experiment. While the model was able to capture the pattern of SRTs across masker types rather well, it systematically underestimated SRTs (overestimated performance) by about 7–10 dB for speech maskers, reflecting the influence of IM that is not captured by the model. This resulted in rather large *rms err* values across masker types that were between 7 and 7.6 dB for co-located sources depending on the cues available, and between 4.8 and 7.9 dB for

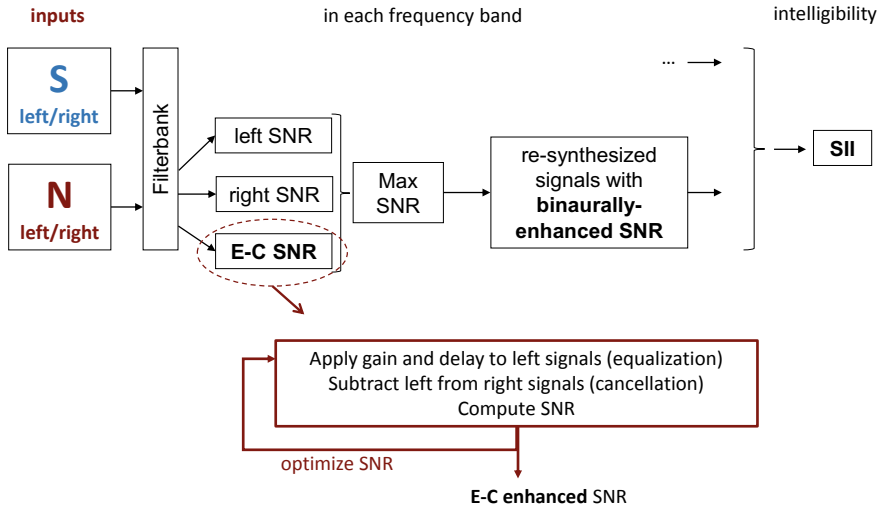


Fig. 1 Schematic of the computational SNR-based model proposed by Beutelmann and Brand (2006). S: signal/target; N: noise/interferer; SNR: signal-to-noise ratio; E-C: equalization-cancellation; SII: speech intelligibility index

spatially separated sources. The *rms err* values computed with the measured and predicted SRMs rather than the SRTs were smaller (between 0.5 and 3.2 dB).

The model of Wan et al. (2010) is conceptually very similar to the model of Beutelmann and Brand (2006), except that the direct implementation of the E-C process uses time-varying jitters in time and amplitude. This model gave accurate predictions for SRTs measured with up to three (noise) interferers at different positions (Hawley et al. 2004). It was only tested in anechoic conditions, but for four types of interferers: stationary and speech-modulated noises, speech and reversed speech. When the free parameter of the model (SII criterion) was fitted only once to the whole data set, *rms err* was 3.6 dB. When the parameter was fitted independently for each interferer type (thus resulting in a different model each time), the *rms err* was 0.7 dB (noise), 1.4 dB (modulated noise), 4.1 dB (speech), and 3.3 dB (reversed speech). Because predicted SRTs were presented only for the scenario using multiple model versions (model parameter changed each time the number and type of interferers varied), it is difficult to tell if a single version of the model could explicitly predict the differences across interferer types. This issue remained for a revised version of the model (Wan et al. 2014), which uses E-C parameters varying across short-time frames along the duration of the stimuli, thus improving the possibility for the model to cancel the dominant masker over time when the direction of the dominant masker varies in time. However, it also uses long-term SNRs calculated over the whole stimulus duration, so that short-time variations in SNR are not explicitly considered, preventing the effect of masker envelope modulations from being explicitly predicted.

Analytical SNR Models

“Analytical” models are binaural models in which the direct implementation of equalization and cancellation is replaced by a predictive equation and the resulting prediction of binaural unmasking is added to a better-ear SNR. One of the differences across analytical models is the equation used for predicting binaural unmasking. In the model proposed by Levitt and Rabiner (1967b), speech intelligibility in noise is predicted from the computation of the AI. Binaural unmasking is taken into account by assuming that the effective SNR in each frequency band is increased by the binaural masking level difference (BMLD) for pure tone detection in noise at the center frequency of the band, using BMLD predictions from Durlach (1963). Predictions based on this model were fairly consistent with previous binaural unmasking data collected by the same authors (Levitt and Rabiner 1967a), in which portions of either the speech or broadband noise spectrum (with varying cutoff frequencies) were subjected to an interaural phase reversal. The maximum absolute difference between the measured and predicted binaural intelligibility level differences produced by phase inversion was 4.1 dB.

Extending this approach, Zurek (1993) proposed a model describing SRM in anechoic situations. Better-ear listening is simulated by computing the SNRs at the two ears by frequency band and taking the better of the left and right SNRs in each band. Binaural unmasking is then taken into account by increasing the better ear SNR by the size of the BMLD in each band, this BMLD being estimated for the given set of interaural parameters using a simplified expression proposed by Colburn (1977). The broadband prediction is computed as the AI-weighted sum of the resulting SNRs. The model predictions were compared with measured SRMs taken from three studies involving a single stationary noise interferer and a frontal target. Across thirteen tested noise azimuths, the maximum absolute difference between measured and predicted SRMs was 4 dB. The comparison of measured and predicted SRMs in six conditions for speech with a 0.5-ms ITD or a phase inversion against a diotic white noise led to a maximum and a mean difference of 3.3 dB and 1.5 dB, respectively. Predictions of the head-shadow advantage were generally larger than the measured effects, whereas binaural-unmasking advantages were predicted fairly well. The models proposed by Levitt and Rabiner (1967b) and Zurek (1993) cannot be applied to reverberant situations, because they do not take into account the interaural coherence of the interferer, which mediates the effect of reverberation on binaural unmasking. Moreover, these models were not tested with non-stationary interferers.

Lavandier and Culling (2010) proposed a model that can be used in reverberant conditions (Fig. 2). Better-ear listening is estimated from the SNR computed as a function of frequency at each ear, selecting band-by-band the ear for which the ratio is the highest. Ratios are weighted according to the SII, and integrated across frequency to provide a broadband better-ear SNR. Binaural unmasking is modeled by increasing the SNR by the size of the BMLD for pure tone detection in noise in each frequency band. BMLDs are estimated from the interaural phase differences of target and interferer and the interaural coherence of the interferer, using an equation proposed by Culling et al. (2004, 2005). Unlike the BMLD estimations used by Levitt

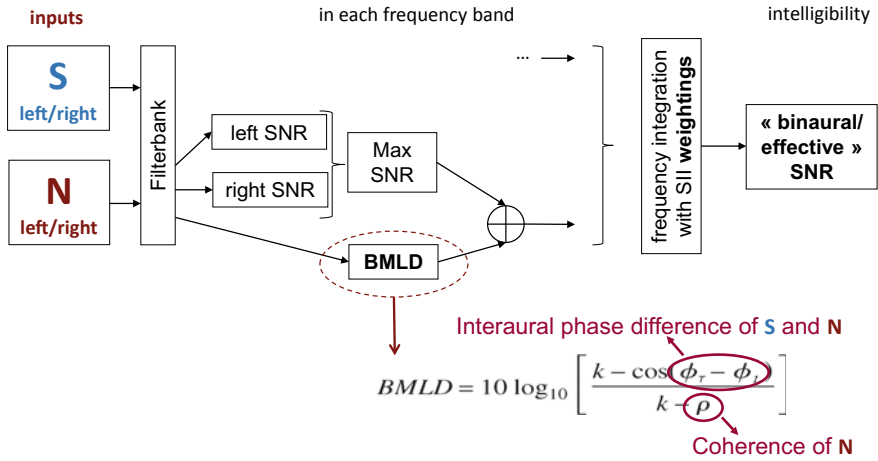


Fig. 2 Schematic of the analytical SNR-based model proposed by Lavandier and Culling (2010). S: signal/target; N: noise/interferer; SNR: signal-to-noise ratio; BMLD: binaural masking level difference; SII: speech intelligibility index

and Rabiner (1967b) and Zurek (1993), this equation can describe the decorrelation of the interferer at the listener’s ears by reverberation and the corresponding impairment of binaural unmasking, allowing predictions in reverberant situations. The BMLD values are then integrated across frequency using the SII weightings to provide a broadband binaural unmasking advantage. The “effective” SNR is obtained by adding the binaural unmasking advantage to the better-ear ratio to predict the overall effect of binaural hearing. For this procedure, it is assumed that the contributions of the two mechanisms are additive.

This model was first used to describe SRTs measured with a single noise interferer simulated at different distances from the listener in virtual rooms varying in size and absorption (Lavandier and Culling 2010). A 0.95–0.97 *corr* was obtained, with *max err* and *mean err* measured below 0.8 dB and 0.3 dB, respectively. Because no head shadow was simulated in this study, the latter essentially validated the binaural unmasking component of the model. In a following investigation, the effect of head shadow was simulated (Lavandier et al. 2012). In addition, stimuli without ITDs but preserved spectral envelope were also used to test the better-ear component of the model specifically. This further validation considered five real rooms and between one to three simultaneous stationary noise interferers, as well as the prediction of the data from Beutelmann and Brand (2006) that involved different rooms, bearings, and distances. *Corr* values ranging from 0.95 to 0.99 were obtained. Across these two studies, the model components were tested both in combination and isolation. They were also validated across a wide range of anechoic conditions (Jelfs et al. 2011; Culling et al. 2013), in situations involving from one to six noise interferers, using SRTs obtained in different laboratories, using different measurement procedures and

different languages (Dutch, English, German). The evaluation produced *corr* values ranging from 0.86 to 0.99.

Following Rhebergen and Versfeld (2005), and Beutelmann et al. (2010), the binaural model of Lavandier and Culling (2010) was adapted to handle non-stationary noises by considering the signals within short-time frames, applying the stationary model within each frame, and averaging the resulting predictions over all frames (Collin and Lavandier 2013). This revised model was used to predict binaural speech intelligibility in the presence of multiple non-stationary noises, varying in modulation depth and spatial location in rooms. Across three experiments, *corr* values between 0.84 and 0.90 were obtained. The *max err* and *mean err* values were measured below 1.6 dB and 0.7 dB, respectively.

3.2 Modulation-Based Models

In contrast to SNR-based models, modulation-based models predict the intelligibility of speech based on the intactness of its temporal modulations. Consequently, these models view speech modulations, rather than speech energy, as the critical factor for intelligibility.

The loss of intelligibility associated with the temporal smearing of target speech by reverberation can be predicted by the Speech Transmission Index (STI; Houtgast and Steeneken 1985). The STI evaluates the loss of amplitude modulation in the speech when it is mixed with multiple delayed versions of itself reflected by room boundaries. The STI also considers the effect of background noise. Wijngaarden and Drullman (2008) developed a binaural version of the STI to predict consonant-vowel-consonant scores in about forty conditions, including speech in quiet and in the presence of a stationary-noise source at different SNRs. The forty conditions included four listening environments: an anechoic room, a listening room, a classroom, and a cathedral. This model offers the advantage of predicting the smearing effect of reverberation on the target speech. However, it also makes the initial assumption that the target is the only source of modulation in the signals reaching the listener's ears. The aim is to look for modulation to identify the position/interaural parameters of the target. This approach does not offer any opportunity for extension to more realistic cases where interferers are modulated noise or speech. Unfortunately, in these cases, modulations are present in both target and interferer, and this cue can no longer be used to distinguish the sound sources. Moreover, since model predictions were only compared to the STI reference curve instead of measuring the goodness of fit to the data, a direct evaluation of performance and comparison with other models is not possible.

Jørgensen and Dau (2011) proposed a monaural model predicting intelligibility based on the modulation spectrum of the sources. The Signal-plus-Noise-to-Noise Ratio, (S+N)NR, is computed from the normalized variance of the envelope fluctuations of the noisy-speech and noise-alone signal inputs within audio-frequency and modulation-frequency bands. An ideal observer fitted to the speech

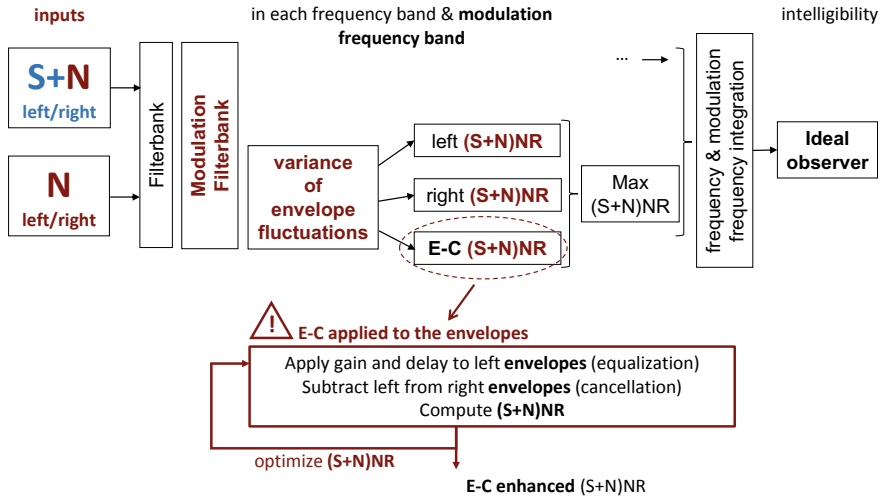


Fig. 3 Schematic of the modulation-based model proposed by Chabot-Leclerc et al. (2016). S: signal/target; N: noise/interferer; (S+N)NR: signal-plus-noise-to-noise ratio; E-C: equalization-cancellation. Note: the model also decomposes the signals in the time domain, using a time-frame duration dependent on the modulation frequency. This decomposition is omitted here for the sake of clarity

material is then used to derive the predicted SRT from the (S+N)NR integrated across audio and modulation frequencies. This model was shown to describe the effects of additive stationary noise, reverberation (temporal smearing), as well as nonlinear speech processing with spectral subtraction (noise reduction). A multi-resolution version of the model was then proposed to predict the increased intelligibility due to dip listening in the presence of fluctuating noise (Jørgensen et al. 2013). Following the idea of applying the stationary model within short-time frames (Rhebergen and Versfeld 2005), the (S+N)NR is computed within frames whose duration depends on the modulation-filter frequency. Accurate predictions were obtained in two conditions with stationary noises (*rms err* of 0.5 dB), five conditions with fluctuating noises (*rms err* of 0.8 dB), five conditions with reverberation (*rms err* of 0.6 dB) and six conditions with spectral subtraction (*rms err* of 1.3 dB vs. 0.5 dB for the stationary model). On the other hand, the model was not able to describe the loss of intelligibility observed for target speech with amplified modulation content (Jørgensen et al. 2015), pointing to a limitation of the model framework that considers any modulation of the speech to be important for intelligibility.

A binaural version of the multi-resolution model has been recently tested to predict SRM (Chabot-Leclerc et al. 2016). It takes the noisy-speech and noise-alone signals at each ear as inputs (Fig. 3). The monaural model of Jørgensen et al. (2013) is applied to the envelopes at the left ear, those at the right ear, and those resulting from the E-C

mechanism proposed by Wan et al. (2014) but applied to the signal envelopes² rather than to the signals themselves. The highest (S+N)NR of the three is selected in each time frame for each modulation and audio frequency, before integration across time and frequencies. Model predictions were compared to anechoic SRTs measured by Hawley et al. (2004) with one to three stationary or fluctuating noises (*corr* of 0.91, *rms err* of 3 dB) and to SRTs measured by Beutelmann et al. (2010) with a single stationary or fluctuating noise and a speech target both placed at various azimuths and distances in three rooms (*corr* of 0.91 and *rms err* of 6.5 dB, vs. 0.89 and 3.6 dB, respectively, for the model of Beutelmann et al. 2010). The model was also applied to a situation in which a varying broadband ITD was applied to the sources (so that only binaural unmasking was involved), and the observed 4-dB SRM was only predicted partially by the model (2-dB SRM). Predictions were on average well-correlated to the data (overall *corr* of 0.91), but this does not reflect the rather large prediction errors occurring in many conditions (with *max err* sometimes reaching 7–10 dB), resulting in large *rms err* values compared to those obtained with the model of Beutelmann and colleagues or during the validation of the monaural version of the model (Jørgensen et al. 2013).

3.3 Correlation-Based Models

The correlation-based models predict intelligibility by looking at the correlation between a clean version of the target speech used as a reference and the noisy and/or processed speech. This type of model was developed in particular to describe the deleterious effect of non-linear speech processing.

The Short-Time Objective Intelligibility (STOI) index is a monaural indicator proposed for predicting the intelligibility of noisy and non-linearly processed speech (Taal et al. 2011). It is computed in two steps. First, a time-frequency decomposition is applied to both the clean and noisy speech, which are assumed to be time-aligned. The frames where the target is silent are removed. These silent moments are detected on the clean speech, but frames are removed from both signals. The correlations between the envelopes of the clean and noisy speech are computed within each time-frequency unit, and then averaged across units; the assumption being that the intelligibility of the noisy/processed speech is related to this average correlation. STOI rates intelligibility on a scale from -1 to 1 . This rating can be mapped to intelligibility ratings in percent correct, or a reference SRT can be chosen to compute predicted SRTs in different conditions.

Andersen et al. (2016) recently proposed a binaural extension of a slightly modified version of STOI (DBSTOI for Deterministic Binaural STOI). The model inputs are the left and right ear signals for the clean and noisy/processed speech (Fig. 4),

²Note that envelopes are extracted here by low-pass filtering the signals below 770 Hz. At low frequencies, where the E-C mechanism is mostly assumed to work, these envelopes still contain a significant amount of the fine structure information required for E-C to take place.

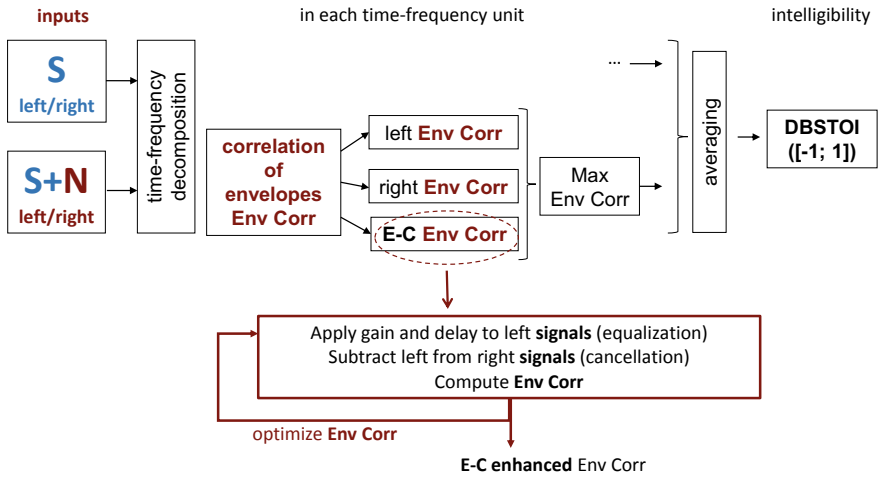


Fig. 4 Schematic of the correlation-based model proposed by Andersen et al. (2016). S: signal/target; N: noise/interferer; Env Corr: correlation of the envelopes; E-C: equalization-cancellation; DBSTOI: deterministic binaural short-time objective intelligibility index

assumed to be time-aligned at each ear. During the time-frequency decomposition, the frames where the target is silent at both ears are removed. A modified E-C stage is introduced before the computation of correlations, to combine the left and right signals into a single clean signal and a single noisy/processed signal, while modeling binaural E-C processing. This stage follows the principle of Beutelmann et al. (2010), in which a direct implementation of an E-C process is used, applying interaural time shifts and gains to the signals and subtracting one from the other. The time-shifts and gains are chosen to maximize the model output, independently for each time frame and frequency band. Here, the modified E-C stage does not maximize the SNR (which cannot be computed because target and interferer are not available individually to the model), but instead it maximizes the correlation between the squared envelopes of the clean and noisy speech. This correlation is also estimated for the signals at each ear individually, as a “better-ear option” corresponding to an infinite gain included in the optimization process. STOI is then calculated on the E-C processed “monaural” clean and noisy speech. For diotic signals, DBSTOI is equivalent to STOI.

Andersen et al. (2016) showed that DBSTOI could predict the effect of nonlinear (diotic) speech processing with the same accuracy as STOI (0.96 correlation between measured and predicted percent correct in the presence of different stationary and non-stationary noises). It predicted the SRM for a frontal target masked by a single stationary noise simulated at different azimuths in anechoic space with the same accuracy as two existing binaural models. A *corr* value of 0.99 was obtained for the three models. The prediction errors were measured at 0.5 dB versus 0.4 dB for the model of Beutelmann et al. (2010) and 0.5 dB for the model of Jelfs et al. (2011). A metric very close to *rms err* was used for this measurement. DBSTOI could also pre-

dict the effects of nonlinear speech processing and SRM taking place simultaneously in the presence of a stationary noise, but underestimated SRTs (overestimated intelligibility) by 2–3 dB for non-stationary noise. Still in anechoic conditions, the model could predict the effect of beamforming, but generally failed in the presence of two noises. This could indicate that the model might require a different mapping between correlation and intelligibility (or reference SRT) depending on the complexity of the situation (e.g., type or number of interferers).

3.4 *Combination Models*

Some models have been proposed that combine two existing models in order to benefit from the advantages of each original model while overcoming some of their limitations to extend their scope of application.

Apart from the STI, another monaural indicator that can be used to take into account the temporal smearing of target speech by reverberation is the Useful-to-Detrimental (U/D) ratio. This SNR-based indicator regards the early reflections of the target as useful and as the “signal” because they reinforce the direct sound (Bradley et al. 2003), whereas the late reflections are regarded as detrimental and effectively a part of the noise (Lochner and Burger 1964; Bradley 1986; Bradley et al. 1999). This indicator has been combined with binaural models, as described below. It is important to mention that the separation of the reflections into useful and detrimental often used the equivalent of a rectangular temporal window with the early/late limit as the single parameter and its value changed quite significantly across studies. An early/late limit of 50 ms has been used very commonly (Roman and Woodruff 2013; Arweiler and Buchholz 2011; Bradley et al. 2003; Soulodre et al. 1989), but other studies also used a limit of 35 ms (Bradley 1986), 80 ms (Bradley 1986), and 100 ms (Lochner and Burger 1964; Rennies et al. 2011, 2014).

To be able to make SRM predictions also for reverberated targets (more or less smeared by reverberation), Rennies et al. (2011) extended the binaural model of Beutelmann and Brand (2006) using three alternatives: the modulation transfer function (MTF) also used in the STI, the definition (D50, ratio of early-to-total impulse response energy; ISO 3382 1997), and the U/D ratio. In the first two approaches, SRM and temporal smearing are processed separately: the SNRs obtained with the binaural model applied to the entire speech and noise signals are corrected a posteriori by either measuring the MTF or D50 of the Binaural Room Impulse Response (BRIR) corresponding to the target. In the third approach, this impulse response is split into early and late parts that are convolved with the speech signal to create an “early speech” signal and a “late speech” signal. The prediction process is then similar to that of Beutelmann and Brand (2006) except that the original target signal is replaced by the early speech and the late speech is added to the interferer, so that the detrimental influence of late reflections is taken into account before the binaural process.

Using data measured in a virtual room with a stationary noise interferer and a target tested at four distances from the listener and three azimuth separations, the D50 and U/D versions proved to be equivalent (*corr* of 0.98 and *rms err* below 1.4 dB) and better than the MTF version (*corr* of 0.93 and *rms err* of 3 dB). Rennie et al. (2014) further tested the three modeling approaches on the data of Warzybok et al. (2013) that involved a frontal target smeared by a single reflection (varying in delay and azimuth) in the presence of an anechoic frontal, diffuse, or lateral noise interferer. On this particular data set, the U/D approach proved to be the most suitable to describe SRM and temporal smearing of speech simultaneously, providing a *corr* value of 0.97 and an *rms err* of 0.9 dB across sixty-two conditions.

Leclère et al. (2015) proposed a different model to simultaneously account for temporal smearing, SRM, and binaural de-reverberation in reverberant environments. It combines the binaural model of Lavandier and Culling (2010) predicting SRM of a near-field target from multiple stationary noise interferers and a U/D decomposition taking into account the temporal smearing effect of reverberation on speech transmission. The target BRIR is first separated into an early and a late part. The early part constitutes the useful component. The late part is combined with the BRIRs of the interferers to form the detrimental component. These BRIRs are concatenated rather than added to preserve phase information and avoid constructive/destructive interference (Jelfs et al. 2011; Lavandier et al. 2012). The binaural model is then applied to the useful and detrimental components in the same way as it was previously applied to the target and interferer BRIRs. The influence of the early/late separation (temporal window shape and limit values) used in the model was investigated systematically.

Model predictions were compared to SRTs measured in three experiments from the literature (Rennie et al. 2011; Lavandier and Culling 2008), involving realistic reverberation from different rooms, SRM from a stationary noise interferer, target smearing, and binaural de-reverberation. Two versions of the model were tested: a room-dependent model for which the parameters were adjusted in each room and a room-independent model with fixed parameters across rooms. The room-independent model was tested on a fourth data set that involved four rooms not used to define its parameters (Wijngaarden and Drullman 2008). Predictions obtained with the room-dependent model accurately fitted the experimental data (*corr* above 0.90, *max err* and *mean err* below 1.2 dB and 0.7 dB, respectively). The room-independent model was less accurate even though it predicted all trends in the data (*corr* above 0.86, *max err* and *mean err* below 2.1 dB and 1 dB, respectively). In particular, the room-independent model was less accurate to predict the data of Wijngaarden and Drullman (*corr* of 0.96, *max err* of 4.9 dB and *mean err* of 1.8 dB). The room dependence of the model parameters might indicate an inherent limitation of the approach and could partially explain the wide range of early/late limits encountered in the literature.

Despite its limitations, the model by Leclère et al. (2015) proposes a unified interpretation of perceptual mechanisms usually considered separately in the literature. Temporal smearing and binaural de-reverberation can be interpreted simply in the framework of SRM. Temporal smearing during speech transmission is just masking of the early target (useful) from a particular interferer: the late target (detrimental). The late target is just an additional masker, treated like any other interfering

source by the model. Its effect appears at high levels of reverberation (Lavandier and Culling 2008) because the late target needs to be sufficiently energetic to become a non-negligible new source of interference. According to the model, binaural de-reverberation can be understood simply in terms of SRM of the early target from this particular interferer.

In order to propose SRM predictions when the target speech and/or noise interferer are not available separately, Cosentino et al. (2014) developed a different combination model—a binaural extension of the Speech-to-Reverberation Modulation energy Ratio (SRMR; Falk et al. 2010)—that predicts the SRM in anechoic and reverberant conditions for a frontal speech target and a single stationary noise source. The model takes as inputs the noisy speech mixture at each ear. The SRMR aims to capture alterations in the modulation spectrum produced by noise and reverberation. A ratio is computed between the energy in the low modulation frequencies (below 20 Hz), attributed mostly to speech, to that of high modulation frequencies (between 20 and 128 Hz), mostly attributed to noise and reverberation. The better-ear component is estimated by taking the best of the left and right ear monaural SRMRs. The mapping of the SRMR to a better-ear advantage in decibel is obtained by fitting the model to the anechoic SRM data. The binaural unmasking component is evaluated using the same BMLD equation as the model of Lavandier and Culling (2010). The interaural phase of the frontal target is known. The interaural cross-correlation function of the noisy speech is computed, and the highest coherence value having a phase different from that of the target is assigned to the noise with its corresponding phase. The two model components are computed across time frames (using two different temporal resolutions) and audio-frequency bands, then averaged across time and integrated across frequency using the SII weightings. The model was evaluated using SRTs measured in anechoic space and in a classroom for a noise placed at nine azimuth angles in the frontal hemisphere. SRTs were not directly predicted, so that the temporal smearing of the target in the classroom could not be described, but the correlation between measured and predicted SRMs was 0.93, with a maximum absolute difference of 2.5 dB. It is not apparent how this model could be extended to a speech-modulated interferer or multiple interferers.

3.5 Segregation-Based/Glimpsing Models

Another class of models was developed specifically to deal with speech masked by speech. There have been several attempts over the years to explain various aspects of masked speech perception in terms of local spectro-temporal “glimpses” (e.g., Cooke 2006). The idea here is that speech is spectro-temporally quite sparse, and in many mixtures of sounds, there are epochs in which the local SNR is high (a criterion needs to be defined) and a rather clean representation of the target sound is available, even if the overall SNR is disadvantageous. Intelligibility is then assumed to be correlated with the proportion of target glimpses among the time-frequency units. One much-discussed aspect of this idea is that the listener needs a way to identify

where in the spectro-temporal plane these high-SNR target glimpses are located and to distinguish them from glimpses that are dominated by competing sounds. Recently, a family of models has appeared in the literature that demonstrates how this problem might be solved for binaural mixtures that contain spatial information relating to the competing talkers. To the extent that the spatial cues enable a listener to sort the acoustic mixture appropriately, they can be viewed as releasing IM. Thus, unlike the energy-based models described above, these models can account for at least some effects of IM. It is worth noting that this general approach also appears in the field of computational scene analysis (e.g., Roman et al. 2003; Srinivasan and Wang 2008) with the primary goal of improving automatic speech recognition rather than predicting human behavior.

The model proposed by Mi and Colburn (2016) applies E-C processing to each time-frequency unit in the stimulus, with the equalization parameters set according to the known direction of the target. In a twist on the usual E-C process, energy from the target direction is canceled, and the amount of canceled energy is used as an indicator of how dominant the target was in that unit. This provides a means for selecting time-frequency units that are dominated by the target. Segregation is then implemented by retaining these target-dominated tiles and eliminating the complementary masker-dominated tiles according to a binary mask. The model is combined with the coherence-based SII (Kates and Arehart 2005) to predict SRM for a given set of stimuli. The model was evaluated using data from an experiment in which a frontal speech target was masked by two speech maskers located symmetrically to either side (Marrone et al. 2008). The model was able to predict measured SRTs reasonably well, albeit with over-predictions of around 3–4 dB that were interpreted in terms of the lack of noise in the E-C process or the optimal selection of time-frequency units.

Josupeit and Hohmann (2017) proposed a model for analyzing multitalker mixtures based on glimpses, motivated by the idea that the glimpses contain useful and robust information for localization, talker identification, and word identification. Their model first identifies robust glimpses in the mixture (which could belong to the target or the maskers) based on the strength of periodicity, and then classifies the glimpses as target or masker based on comparisons to templates. The model was evaluated on stimuli from the study by Brungart and Simpson (2007), who measured word recognition in a closed-set listening task using spatialized competing talkers. The task was to recognize the color and number words of the talker who uttered a specific call-sign. The model was able to identify keywords at a comparable rate to human subjects correctly. However, the model requires extensive a priori knowledge in the form of a precise set of clean templates for all possible talkers, locations, and keywords.

Tang et al. (2016) developed a binaural distortion-weighted glimpse proportion metric, which can be correlated with intelligibility scores in anechoic conditions (Fig. 5). It can be computed from two alternative input forms: either individual binaural recordings of target and masker at the ears or monophonic recordings from each sound source along with their locations (azimuth and distance) used to estimate the binaural recordings. The metric evaluates the proportion of target glimpses among the time-frequency units constituting the mixture. This proportion is weighted in each

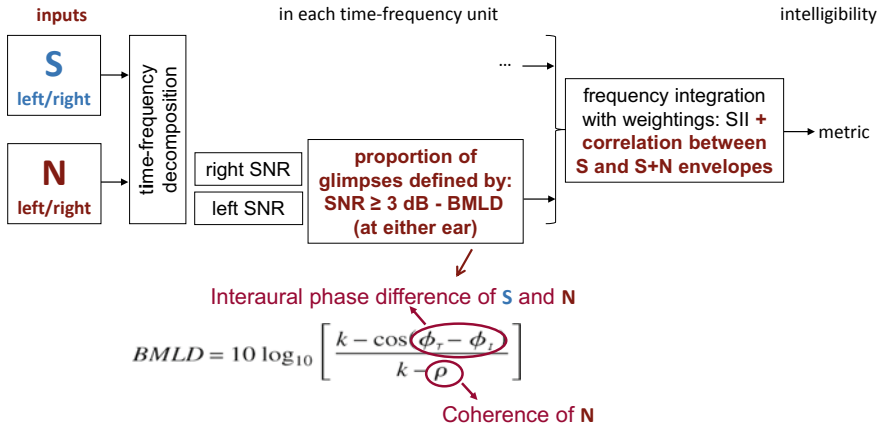


Fig. 5 Schematic of the glimpsing model proposed by Tang et al. (2016). S: signal/target; N: noise/interferer; SNR: signal-to-noise ratio; BMLD: binaural masking level difference; SII: speech intelligibility index

frequency band by the masker-induced distortion of the speech envelope (obtained by computing the cross-correlation between the envelopes of the clean speech and of the speech+masker mixture) and then integrated across frequency using the SII weightings. A glimpse is defined as a unit in which the target level plus the BMLD is 3 dB above the masker level so that binaural unmasking is taken into account at this stage. The BMLD is estimated using the same equation as in the model of Lavandier and Culling (2010). To take into account better-ear listening, glimpses are computed separately for the left and right ears and combined to produce binaural glimpses for all time-frequency units where either or both individual ears produce a glimpse. The resulting metric was correlated to intelligibility scores measured with noise and speech maskers mixed with the target at two SNRs. In a first experiment, the frontal target was presented with a single masker simulated at different azimuths and distances. In a second experiment, the target and two or three maskers were presented at different azimuths in twelve spatial configurations. Overall, the metric was well-correlated to the listener scores, with correlations of 0.95 and 0.91 for single- and multi-masker conditions, respectively; indicating that the metric followed the trends in the data well, even if the magnitude of the effects was smaller in the metric predictions, particularly for the multi-masker experiment. The mapping between the metric and intelligibility scores remains to be estimated. It could be dependent on the specific conditions and masker types considered.

3.6 Multiple Regression Models

Following on a descriptive model proposed by Bronkhorst (2000), Jones and Litovsky (2011) predicted the amount of SRM in several two-interferer configurations from five different studies, using multiple regression models based on two contributions, namely the asymmetry of the two-interferer configuration and the angular separation of target and interferers. Although only a frontal target can be considered, separate model versions could be used for speech and noise interferers. Good fits between measured and predicted SRMs were reported for both noise maskers (*corr* of 0.93, *max err* of 2.5 dB and *mean err* of 0.8 dB) and speech maskers (*corr* of 0.94, *max err* of 2.4 dB and *mean err* of 0.7 dB). However, these performance statistics are misleading because they do not reflect the performance of a single model. They were computed by aggregating the predictions obtained with several versions of the model, corresponding to multiple regressions fitted independently for each data set, and in the case of the speech models, for each number of maskers within each data set. Also, the model versions were tested only on the data sets used to define their fitted parameters, resulting in descriptive models whose predictive power outside these data sets is unknown. Since the models are designed to predict SRMs between spatial configurations rather than the SRTs in these configurations for different types of interferers, they cannot predict the effect of envelope modulations across interferer types. For example, the noise model cannot directly predict the dip-listening advantages associated with noises having different depths of modulation in a given configuration. The models were tested on data measured with all sources at the same distance in a small sound-treated room. This room was not anechoic, but reverberation was not varied during the experiment.

4 Usability of the Models in Practice

4.1 Model Inputs

As highlighted above, the available binaural models use different inputs to make intelligibility predictions. The availability of a model's specific inputs will, of course, determine its usability in practice for any given application. The nature of these inputs can be seen as a priori information/knowledge required by the models. Any model computing an SNR (e.g., SNR-based models, but also glimpsing models) needs to access the target (S) and masker (N) signals independently at each ear (Lavandier and Culling 2010; Beutelmann and Brand 2006; Wan et al. 2010). The modulation-based model uses the noise+speech (S+N) and noise-alone (N) signals at the ears (Chabot-Leclerc et al. 2016), but it is not clear whether the realization of the noise needs to be identical for N and S+N, in which case it would be equivalent to requiring S and N separately at the ears.

The model proposed by Cosentino et al. (2014) uses the S+N mixtures directly, without requiring one or the two signals individually at the ears, which could generalize its usability. However, in return, it needs to assume that the target is in front and that there is a single interferer so that in practice, the use of the model is limited to these configurations. The metric developed by Tang et al. (2016) operates with either the target and masker signals independently at each ear or single-channel versions of these signals together with the source locations that need to be known. In the latter case, the signals and locations are used to estimate the binaural recordings, assuming anechoic conditions to estimate ITDs and ILDs. This method cannot be easily generalized to reverberant conditions.

The correlation-based model requires time-aligned noise+speech and clean speech signals as inputs (Andersen et al. 2016). Time alignment is not always available, except with dedicated recordings, and it might not be easy to obtain a posteriori using non-synchronized signals—especially in reverberation or at low SNRs when realizing this alignment is not trivial. The output of the model could well be very dependent on potential errors in the time alignment, which would restrict the applicability of the model.

In terms of other a priori information required, the combination models based on the U/D approach need access to the target BRIR so that useful and detrimental room reflections can be separated (Rennies et al. 2011; Leclère et al. 2015). The segregation-based model of Mi and Colburn (2016) requires knowledge of the direction of the target signal. Finally, while the segregation-based model of Josupeit and Hohmann (2017) does not require specific information about the target or maskers, it does require detailed knowledge about the set of possible utterances, talkers, and locations, and thus, is only suitable for a closed-set task that involves some limited number of these possibilities.

Most of these models do not need the same signal/sentences as used to collect the data. They can use generic signals as long as those signals have precisely the same statistics (in terms of frequency spectrum levels, modulation spectrum, interaural statistics, etc.) as those involved in the listening test to be described or in the new situations to be predicted. Some segregation-based models, however, have only been evaluated using specific speech mixtures and their associated data sets; it is not yet clear whether they can be generalized to other signals and data sets.

4.2 *Speech Material Used for the Target*

The models involving SII calculations (e.g., Beutelmann and Brand 2006; Wan et al. 2010) or STOI (Andersen et al. 2016) require a form of a priori knowledge regarding the speech material for which intelligibility needs to be predicted. This might include the language, the syntactic and linguistic contents, and predictability. Speech material varying in these dimensions will require a different mapping function between the SII/STOI and percent word correct (Kryter 1962; ANSI S3.5 1997) in order to generate absolute predictions of behavioral performance in human listeners. In the-

ory, the mapping could be changed to predict data for other kinds of material, but this has rarely been explored for the models described above. It is also worth noting that many models can only provide *relative* predictions, using a reference condition to fit data and predictions—e.g., the SRT in a co-located condition as in Zurek (1993) and Andersen et al. (2016), or the average SRT across tested conditions as in Lavandier and Culling (2010) and Jelfs et al. (2011), or predicting SRMs rather than SRTs (Cosentino et al. 2014). The modulation-based model of Chabot-Leclerc et al. (2016) involves an ideal observer fitted to the speech material using the SRT obtained in a reference condition.

4.3 Structure of the Competing Maskers

With fluctuating maskers, Collin and Lavandier (2013) showed that listeners can take advantage of the predictability of the timing of the gaps in the maskers. This indicates that the dip-listening advantage might have been overestimated in studies using predictable modulations such as periodic modulations or “frozen” speech modulations across the adaptive SRT measurement. In situations where the interferer gaps are less predictable, such as with competing talkers, then the listener might not be able to use glimpsing optimally, due to the uncertainty in the gap position within the masking sound. This effect of predictability/uncertainty is not explicitly taken into account in current models.

To the authors’ knowledge, none of the binaural models presented above were tested for periodic interferers, or they were tested for speech and noise interferers but predictions across interferer types could not be made because the model was changed. Given that the magnitude of SRM might be reduced when there is the possibility for F0-based release from masking, this seems to be an important test of the models if they are intended to explain performance in realistic speech mixtures where multiple cues will often be available.

4.4 Room Adaptation for the U/D Models

In the binaural U/D model proposed by Leclère et al. (2015), the best model performance was achieved by adjusting the early/late separation of the room reflections for each tested room (e.g., room dependency of the distinction between the early/useful and late/detrimental reflections for the target). Room-independent parameters did not lead to similar performances, suggesting that a fixed early/late separation might not be sufficient to predict speech intelligibility in rooms, jeopardizing the generalization of the U/D approach to making a priori predictions in any room. Alternatively, the value of the early/late limit in a given room could be seen as additional a priori information required by the models. One might be able to overcome this limitation by modeling other perceptual mechanisms and/or cognitive factors so that predic-

tions can be made in arbitrary rooms—without this a priori knowledge. For instance, previous studies showed that listeners are able to adapt to room acoustics given prior exposure (Watkins 2005; Brandewie and Zahorik 2010). The determination of the early/late parameters in a particular room might be dependent on such adaptation.

5 Incorporating Cognitive Factors

5.1 *Informational Masking and Effects of Attention*

The models discussed above that only take the energy—or the modulation energy—of the incoming target and masker signals into account cannot account for any masking effects that are driven by non-energetic factors. As a result, these models tend to fail when tested with speech maskers that generate substantial IM. However, by quantifying the EM component precisely for various configurations, these models could allow one to estimate the “residual” masking that can be attributed to IM. These estimates are useful for demonstrating the limitations of energy-based models, and also may trigger and inform extensions to existing models in the future. Segregation-based models, on the other hand, explicitly deal with the problem of segregating the target from the maskers, which is one of the key components of IM. As a result, these models do quite well at predicting performance for speech-on-speech tasks (e.g., Mi and Colburn 2016; Josupeit and Hohmann 2017).

Even when competing sounds are spatially separated so that both EM and IM are reduced, the state of a listener’s attention can influence speech understanding. For example, knowledge about the location or timing of an upcoming target, which reduces uncertainty about how to direct one’s attention, can improve intelligibility (Kidd et al. 2005a; Best et al. 2007). In related experiments, Brungart and colleagues showed that randomizing the target voice or location across trials—versus keeping it fixed within a block—can reduce intelligibility scores (Brungart et al. 2001; Brungart and Simpson 2007). It has also been shown that continuity of the location of a target from word to word within a trial is essential for optimal intelligibility in the presence of competitors (Best et al. 2008). These attentional aspects of IM cannot be accounted for by current models of speech intelligibility that only take into account the properties of the stimuli. Extensions of the models might include adding in a “noise” term to deal with failures of attention. However, there is evidence that susceptibility to attentional aspects of IM is highly individualized, being influenced by listener-related factors such as age and musical ability (e.g., Neher et al. 2011; Swaminathan et al. 2015; Clayton et al. 2016).

In situations where uncertainty is relatively low, and attention is focused selectively on the location or voice of a target talker, several studies have demonstrated steady improvements in performance over time, indicative of refinements in selective attention (e.g., Best et al. 2008; Ezzatian et al. 2012). These across-time effects,

which occur in the absence of any change in the stimulus, are certainly not captured by any of the current binaural models, all of which are stimulus-driven.

5.2 *Intelligibility Versus Comprehension*

Speech understanding is generally measured by way of a word or sentence recognition test in which a single utterance is presented and followed by a silent period in which listeners are required to recall the utterance, for example, by speaking it out loud, typing it on a keyboard, selecting words on a touchscreen. However, many researchers over the years have pointed out the many ways in which the requirements of this task differ from those encountered in real-world speech communication. Typically, during a conversation, a listener must follow ongoing speech, rather than discrete, and they must not only recognize the words but comprehend their meaning, and try to determine the intention of their partner. Moreover, they often do other tasks in parallel, such as filling in words that were missed, making predictions about what someone is going to say, or formulating replies. This extra layer of processing means that true speech comprehension, especially in the context of two-way communication, is much more cognitively demanding than word or sentence recall.

Several speech-comprehension tests have been developed to try to capture some of these real-world aspects (e.g., Best et al. 2018; Xia et al. 2017). The test proposed by Best et al. (2018) was closely compared to a sentence test, and it was found that cognitive factors were more strongly associated with comprehension scores than with sentence scores. Specifically, comprehension scores were better than sentence scores for listeners with strong cognitive skills, but the opposite was true for listeners with weak cognitive skills. Thus sentence scores may not adequately capture the real-world communication abilities of a given listener. If binaural speech intelligibility models are to predict speech understanding in realistic situations for specific listeners or groups of listeners, it seems that cognitive factors would need to be accounted for. Comprehension scores also appear to be less sensitive to changes in SNR than are sentence scores, which is an issue for current models. Almost nothing is currently known about the interaction between task (sentence recognition vs. speech comprehension) and spatial separation, which would have particular implications for binaural speech intelligibility models. This is an avenue that deserves further attention.

6 Conclusion

The successive versions of the given model were developed to deal with more and more complex situations and overcome the limitations of the previous versions. Different, independent model categories have been proposed to address different types of key information for intelligibility, e.g., modulation or energy. It could also

be sometimes the case that a different model is developed because its inputs better fit the researchers' needs—e.g., the case that 'S' and 'N' are not available separately. Some approaches could also be particularly suited to a given type of application (e.g., non-linear speech processing). One clear benefit of relying on different assumptions is that these modeling approaches are complementary.

SNR-based models can accurately predict SRM and can deal with modulated noises or reverberated targets. The SNR approach is also very suitable for modeling audibility in terms of internal noise so that predictions can be proposed for hearing-impaired listeners (Beutelmann and Brand 2006; Lavandier et al. 2018). On the other hand, computing the SNR requires access to the target and masker signals independently. The modulation models offer an interesting framework and can predict some effects of non-linear speech processing; however, they have not been fully explored in binaural conditions. The correlation model of Andersen et al. (2016) predicted very well the anechoic SRM for a single stationary noise as well as some effects of non-linear speech processing. However, the requirement for time-aligned signals might compromise its use in reverberation. It is not clear whether the mapping between correlation (STOI) and intelligibility could be dependent on the specific conditions tested, such as the type or number of maskers. The glimpsing model of Tang et al. (2016) is interesting in that it incorporates some aspects of three modeling approaches: SNR, modulation, and correlation.

All these models have common limitations that will pave the way for further research. None can predict SRM for speech maskers in a way that accounts for the strength of IM present. None can describe the release from EM and IM afforded by differences in F0. Only two models have been extended for hearing-impaired listeners (Beutelmann and Brand 2006; Lavandier et al. 2018), whereas two different models can describe the effect of non-linear speech processing as found in hearing aids (Jørgensen et al. 2013; Andersen et al. 2016). It seems that a combination of approaches is needed in order to predict binaural speech intelligibility in the real world for all kinds of listeners and situations.

Acknowledgements Mathieu Lavandier is taking part in the Labex CeLyA funded by the French National Research Agency (ANR-10-LABX-0060/ANR-16-IDEX-0005), while Virginia Best is supported by the U.S. National Institutes of Health (NIH-NIDCD award DC015760). Thank is due to two anonymous experts for peer-reviewing an earlier manuscript of this chapter.

References

- Andersen, A.H., Z.-H. Tan, J.M. de Haan, and J. Jensen. 2016. Predicting the intelligibility of noisy and nonlinearly processed binaural speech. *IEEE Transactions on Audio, Speech, Language Process* 24 (11): 1908–1920.
- ANSI S3.5. 1997. Methods for calculation of the speech intelligibility index. New York: American National Standards Institute.
- Arbogast, T.L., C.R. Mason, and G. Kidd. 2002. The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America* 112: 2086–2098.

- Arweiler, I., and J.M. Buchholz. 2011. The influence of spectral characteristics of early reflections on speech intelligibility. *The Journal of the Acoustical Society of America* 130 (2): 996–1005.
- Best, V., G. Keidser, K. Freeston, and J.M. Buchholz. 2018. Evaluation of the NAL dynamic conversations test in older listeners with hearing loss. *International Journal of Audiology* 57: 221–229.
- Best, V., C.R. Mason, and G. Kidd. 2011. Spatial release from masking in normally hearing and hearing-impaired listeners as a function of the temporal overlap of competing talkers. *The Journal of the Acoustical Society of America* 129: 1616–1625.
- Best, V., C.R. Mason, E.R. Thompson, and G. Kidd. 2013. An energetic limit on spatial release from masking. *Journal of the Association for Research in Otolaryngology* 14: 603–610.
- Best, V., E. Ozmeral, and B.G. Shinn-Cunningham. 2007. Visually-guided attention enhances target identification in a complex auditory scene. *Journal of the Association for Research in Otolaryngology* 8: 294–304.
- Best, V., E.J. Ozmeral, and N. Kopčo, and B.G. Shinn-Cunningham. 2008. Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences* 105: 13173–13177.
- Beutelmann, R., and T. Brand. 2006. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America* 120 (1): 331–342.
- Beutelmann, R., T. Brand, and B. Kollmeier. 2010. Revision, extension, and evaluation of a binaural speech intelligibility model. *The Journal of the Acoustical Society of America* 127 (4): 2479–2497.
- Binns, C., and J.F. Culling. 2007. The role of fundamental frequency contours in the perception of speech against interfering speech. *The Journal of the Acoustical Society of America* 122 (3): 1765–1776.
- Bradley, J.S. 1986. Predictors of speech intelligibility in rooms. *The Journal of the Acoustical Society of America* 80 (3): 837–845.
- Bradley, J.S., R.D. Reich, and S.G. Norcross. 1999. On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility. *The Journal of the Acoustical Society of America* 106 (4): 1820–1828.
- Bradley, J.S., H. Sato, and M. Picard. 2003. On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America* 113 (6): 3233–3244.
- Brandewie, E., and P. Zahorik. 2010. Prior listening in rooms improves speech intelligibility. *The Journal of the Acoustical Society of America* 128 (1): 291–299.
- Brox, J.P.L., and S.G. Nooteboom. 1982. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics* 10: 23–36.
- Bronkhorst, A.W. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica* 86 (1): 117–128.
- Bronkhorst, A.W., and R. Plomp. 1988. The effect of head-induced interaural time and level differences on speech intelligibility in noise. *The Journal of the Acoustical Society of America* 83 (4): 1508–1516.
- Bronkhorst, A.W., and R. Plomp. 1990. A clinical test for the assessment of binaural speech perception in noise. *Audiology* 29: 275–285.
- Bronkhorst, A.W., and R. Plomp. 1992. Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *The Journal of the Acoustical Society of America* 92 (6): 3132–3139.
- Brungart, D.S. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America* 109: 1101–1109.
- Brungart, D.S., and N. Iyer. 2012. Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *The Journal of the Acoustical Society of America* 132 (4): 2545–2556.
- Brungart, D.S., and B.D. Simpson. 2002. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. *The Journal of the Acoustical Society of America* 112 (2): 664–676.

- Brungart, D.S., and B.D. Simpson. 2007. Cocktail party listening in a dynamic multitalker environment. *Perception and Psychophysics* 69: 79–91.
- Brungart, D.S., B.D. Simpson, M.A. Ericson, and K.R. Scott. 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America* 110: 2527–2538.
- Chabot-Leclerc, A., E.N. MacDonald, and T. Dau. 2016. Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain. *The Journal of the Acoustical Society of America* 140 (1): 192–205.
- Clayton, K.K., J. Swaminathan, A. Yazdanbakhsh, J. Zuk, A. Patel, and G. Kidd. 2016. Executive function, visual attention and the cocktail party problem in musicians and non-musicians. *PLoS One* e0157638.
- Colburn, H.S. 1977. Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise. *The Journal of the Acoustical Society of America* 61 (2): 525–533.
- Collin, B., and M. Lavandier. 2013. Binaural speech intelligibility in rooms with variations in spatial location of sources and modulation depth of noise interferers. *The Journal of the Acoustical Society of America* 134 (2): 1146–1159.
- Cooke, M. 2006. A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America* 119 (3): 1562–1573.
- Cosentino, S., T. Marquardt, D. McAlpine, J.F. Culling, and T.H. Falk. 2014. A model that predicts the binaural advantage to speech intelligibility from the mixed target and interferer signals. *The Journal of the Acoustical Society of America* 135 (2): 796–807.
- Culling, J.F., M.L. Hawley, and R.Y. Litovsky. 2004. The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *The Journal of the Acoustical Society of America* 116 (2): 1057–1065.
- Culling, J.F., M.L. Hawley, and R.Y. Litovsky. 2005. Erratum: The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *The Journal of the Acoustical Society of America* 118 (1): 552.
- Culling, J.F., K.I. Hodder, and C.Y. Toh. 2003. Effects of reverberation on perceptual segregation of competing voices. *The Journal of the Acoustical Society of America* 114 (5): 2871–2876.
- Culling, J.F., M. Lavandier, and S. Jelfs. 2013. Predicting binaural speech intelligibility in architectural acoustics. In *The Technology of Binaural Listening*, ed. J. Blauert, 427–447. Berlin-Heidelberg-New York, NY: Springer.
- Culling, J.F., and E.R. Mansell. 2013. Speech intelligibility among modulated and spatially distributed noise sources. *The Journal of the Acoustical Society of America* 133 (4): 2254–2261.
- Culling, J.F., and M.A. Stone. 2017. Energetic masking and masking release. In *The Auditory System at the Cocktail Party*, vol. 60, ed. J. Middlebrooks, J. Simon, A.N. Popper, and R.R. Fay, 41–73. Cham: Springer Handbook of Auditory Research, Springer.
- Culling, J.F., Q. Summerfield, and D.H. Marshall. 1994. Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels. *Speech Communication* 14: 71–96.
- de Cheveigné, A., S. McAdams, J. Laroche, and M. Rosenberg. 1995. Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement. *The Journal of the Acoustical Society of America* 97 (6): 3736–3748.
- Deroche, M.L.D., and J.F. Culling. 2011. Voice segregation by difference in fundamental frequency: Evidence for harmonic cancellation. *The Journal of the Acoustical Society of America* 130 (5): 2855–2865.
- Deroche, M.L.D., J.F. Culling, M. Chatterjee, and C.J. Limb. 2014. Roles of the target and masker fundamental frequencies in voice segregation. *The Journal of the Acoustical Society of America* 136 (3): 1225–1236.
- Durlach, N.I. 1963. Equalization and cancellation theory of binaural masking-level differences. *The Journal of the Acoustical Society of America* 35 (8): 1206–1218.
- Durlach, N.I. 1972. Binaural signal detection: Equalization and cancellation theory. In *Foundations of Modern Auditory Theory*, vol. II, ed. J. Tobias, 371–462. New York: Academic.

- Ewert, S.D., W. Schubotz, T. Brand, and B. Kollmeier. 2017. Binaural masking release in symmetric listening conditions with spectro-temporally modulated maskers. *The Journal of the Acoustical Society of America* 142: 12–28.
- Ezzatian, P., L. Li, K. Pichora-Fuller, and B. Schneider. 2012. The effect of energetic and informational masking on the time-course of stream segregation: Evidence that streaming depends on vocal fine structure cues. *Language and Cognitive Processes* 27: 1056–1088.
- Falk, T., C. Zheng, and W. Chan. 2010. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing* 18: 1766–1774.
- Festen, J.M., and R. Plomp. 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America* 88 (4): 1725–1736.
- Freyman, R.L., U. Balakrishnan, and K.S. Helfer. 2001. Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America* 109 (5): 2112–2122.
- Freyman, R.L., K.S. Helfer, D.D. McCall, and R.K. Clifton. 1999. The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America* 106 (6): 3578–3588.
- George, E.L.J., J.M. Festen, and T. Houtgast. 2008. The combined effects of reverberation and nonstationary noise on sentence intelligibility. *The Journal of the Acoustical Society of America* 124 (2): 1269–1277.
- Glyde, H., J.M. Buchholz, H. Dillon, V. Best, L. Hickson, and S. Cameron. 2013a. The effect of better-ear glimpsing on spatial release from masking. *The Journal of the Acoustical Society of America* 134: 2937–2945.
- Glyde, H., J.M. Buchholz, H. Dillon, S. Cameron, and L. Hickson. 2013b. The importance of interaural time differences and level differences in spatial release from masking. *Journal of the Acoustical Society of America (Express Letters)* 134: EL147–152.
- Hawley, M.L., R.Y. Litovsky, and J.F. Culling. 2004. The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. *The Journal of the Acoustical Society of America* 115 (2): 833–843.
- Houtgast, T., and H.J.M. Steeneken. 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America* 77 (3): 1069–1077.
- ISO 3382. 1997. *Acoustics—Measurement of the reverberation time of rooms with reference to other acoustical parameters*. Geneva: International Organization for Standardization.
- Jelfs, S., J.F. Culling, and M. Lavandier. 2011. Revision and validation of a binaural model for speech intelligibility in noise. *Hearing Research* 275: 96–104.
- Johnsrude, I.S., A. Mackey, H. Hakyemez, E. Alexander, H.P. Trang, and R.P. Carlyon. 2013. Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science* 24 (10): 1995–2004.
- Jones, G.L., and R.Y. Litovsky. 2011. A cocktail party model of spatial release from masking by both noise and speech interferers. *The Journal of the Acoustical Society of America* 130 (3): 1463–1474.
- Jørgensen, S., and T. Dau. 2011. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America* 130 (3): 1475–1487.
- Jørgensen, S., R. Decorsière, and T. Dau. 2015. Effects of manipulating the signal-to-noise envelope power ratio on speech intelligibility. *The Journal of the Acoustical Society of America* 137 (3): 1401–1410.
- Jørgensen, S., S.D. Ewert, and T. Dau. 2013. A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America* 134 (1): 436–446.
- Josupeit, A., and V. Hohmann. 2017. Modeling speech localization, talker identification, and word recognition in a multi-talker setting. *The Journal of the Acoustical Society of America* 142: 35–54.

- Kates, J.M., and K.H. Arehart. 2005. Coherence and the speech intelligibility index. *The Journal of the Acoustical Society of America* 117: 2224–2237.
- Kidd, G., and H.S. Colburn. 2017. Informational masking in speech recognition. In *The Auditory System at the Cocktail Party*, vol. 60, ed. J. Middlebrooks, J. Simon, A.N. Popper, and R.R. Fay, 75–109. Cham: Springer Handbook of Auditory Research, Springer.
- Kidd, G., T.L. Arbogast, C.R. Mason, and F.J. Gallun. 2005a. The advantage of knowing where to listen. *The Journal of the Acoustical Society of America* 118: 3804–3815.
- Kidd, G., C.R. Mason, A. Brughera, and W.M. Hartmann. 2005b. The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acta Acustica United with Acustica* 91 (3): 526–535.
- Kryter, K.D. 1962. Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America* 34 (11): 1689–1697.
- Lavandier, M., J.M. Buchholz, and B. Rana. 2018. A binaural model predicting speech intelligibility in the presence of stationary noise and noise-vocoded speech interferers for normal-hearing and hearing-impaired listeners. *Acta Acustica United with Acustica* 104 (5): 909–913.
- Lavandier, M., and J.F. Culling. 2008. Speech segregation in rooms: Monaural, binaural, and interacting effects of reverberation on target and interferer. *The Journal of the Acoustical Society of America* 123 (4): 2237–2248.
- Lavandier, M., and J.F. Culling. 2010. Prediction of binaural speech intelligibility against noise in rooms. *The Journal of the Acoustical Society of America* 127 (1): 387–399.
- Lavandier, M., S. Jelfs, J.F. Culling, A.J. Watkins, A.P. Raimond, and S.J. Makin. 2012. Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. *The Journal of the Acoustical Society of America* 131 (1): 218–231.
- Leclère, T., M. Lavandier, and J.F. Culling. 2015. Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking and binaural de-reverberation. *The Journal of the Acoustical Society of America* 137 (6): 3335–3345.
- Leclère, T., M. Lavandier, and M.L.D. Deroche. 2017. The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location. *Hearing Research* 350: 1–10.
- Levitt, H., and L.R. Rabiner. 1967a. Binaural release from masking for speech and gain in intelligibility. *The Journal of the Acoustical Society of America* 42 (3): 601–608.
- Levitt, H., and L.R. Rabiner. 1967b. Predicting binaural gain in intelligibility and release from masking for speech. *The Journal of the Acoustical Society of America* 42 (4): 820–829.
- Licklider, J.C.R. 1948. The influence of interaural phase relations upon masking of speech by white noise. *The Journal of the Acoustical Society of America* 20 (2): 150–159.
- Lingner, A.B.G., L. Wiegrebe, and S.D. Ewert. 2016. Binaural glimpses at the cocktail party? *Journal of the Association for Research in Otolaryngology* 17: 461–473.
- Lochner, J.P.A., and J.F. Burger. 1964. The influence of reflections on auditorium acoustics. *Journal of Sound and Vibration* 1 (4): 426–454.
- Marrone, N., C.R. Mason, and G. Kidd. 2008. Tuning in the spatial dimension: Evidence from a masked speech identification task. *The Journal of the Acoustical Society of America* 124: 1146–1158.
- Martin, R.L., K.I. McAnally, R.S. Bolia, G. Eberle, and D.S. Brungart. 2012. Spatial release from speech-on-speech masking in the median sagittal plane. *The Journal of the Acoustical Society of America* 131 (1): 378–385.
- Mi, J., and H.S. Colburn. 2016. A binaural grouping model for predicting speech intelligibility in multitalker environments. *Trends in Hearing* 20: 1–12.
- Moncur, J.P., and D. Dirks. 1967. Binaural and monaural speech intelligibility in reverberation. *Journal of Speech and Hearing Research* 10: 186–195.
- Nábělek, A.K., and P.K. Robinson. 1982. Monaural and binaural speech perception in reverberation for listeners of various ages. *The Journal of the Acoustical Society of America* 71 (5): 1242–1248.

- Neher, T., S. Laugesen, N. Jensen, and L. Kragelund. 2011. Can basic auditory and cognitive measures predict hearing-impaired listeners' localization and spatial speech recognition abilities? *The Journal of the Acoustical Society of America* 130 (3): 1542–1558.
- Plomp, R. 1976. Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise). *Acustica* 34: 200–211.
- Rennies, J., T. Brand, and B. Kollmeier. 2011. Prediction of the influence of reverberation on binaural speech intelligibility in noise and in quiet. *The Journal of the Acoustical Society of America* 130 (5): 2999–3012.
- Rennies, J., A. Warzybok, T. Brand, and B. Kollmeier. 2014. Modeling the effects of a single reflection on binaural speech intelligibility. *The Journal of the Acoustical Society of America* 135 (3): 1556–1567.
- Rhebergen, K.S., and N.J. Versfeld. 2005. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America* 117 (4): 2181–2192.
- Rhebergen, K.S., N.J. Versfeld, and W.A. Dreschler. 2006. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *The Journal of the Acoustical Society of America* 120 (6): 3988–3997.
- Roman, N., D. Wang, and G.J. Brown. 2003. Speech segregation based on sound localization. *The Journal of the Acoustical Society of America* 114: 2236–2252.
- Roman, N., and J. Woodruff. 2013. Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold. *The Journal of the Acoustical Society of America* 133 (3): 1707–1717.
- Schoenmaker, E., T. Brand, and S. van de Par. 2016. The multiple contributions of interaural differences to improved speech intelligibility in multitalker scenarios. *The Journal of the Acoustical Society of America* 139: 2589–2603.
- Schoenmaker, E., S. Sutojo, and S. van de Par. 2017. Better-ear rating based on glimpsing. *The Journal of the Acoustical Society of America* 142: 1466–1481.
- Soulodre, G.A., N. Popplewell, and J.S. Bradley. 1989. Combined effects of early reflections and background noise on speech intelligibility. *Journal of Sound and Vibration* 135 (1): 123–133.
- Souza, P., N. Gehani, R. Wright, and D. McCloy. 2013. The advantage of knowing the talker. *Journal of the American Academy of Audiology* 24: 689–700.
- Srinivasan, S., and D. Wang. 2008. A model for multitalker speech perception. *The Journal of the Acoustical Society of America* 124: 3213–3224.
- Swaminathan, J., C.R. Mason, T. Streeter, V. Best, G. Kidd, and A. Patel. 2015. Musical training, individual differences and the cocktail party problem. *Scientific Reports* 5: 11628.
- Taal, C.H., R.C. Hendriks, R. Heusdens, and J. Jensen. 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing* 19 (7): 2125–2136.
- Tang, Y., M. Cooke, B.M. Fazenda, and T.J. Cox. 2016. A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers. *The Journal of the Acoustical Society of America* 140 (3): 1858–1870.
- Wan, R., N.I. Durlach, and H.S. Colburn. 2010. Application of an extended equalization-cancellation model to speech intelligibility with spatially distributed maskers. *The Journal of the Acoustical Society of America* 128 (6): 3678–3690.
- Wan, R., N.I. Durlach, and H.S. Colburn. 2014. Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers. *The Journal of the Acoustical Society of America* 136 (2): 768–776.
- Warzybok, A., J. Rennies, T. Brand, S. Doclo, and B. Kollmeier. 2013. Effects of spatial and temporal integration of a single early reflection on speech intelligibility. *The Journal of the Acoustical Society of America* 133 (1): 269–282.
- Watkins, A.J. 2005. Perceptual compensation for effects of reverberation in speech identification. *The Journal of the Acoustical Society of America* 118 (1): 249–262.

- Wijngaarden, S.J., and R. Drullman. 2008. Binaural intelligibility prediction based on the speech transmission index. *The Journal of the Acoustical Society of America* 123 (6): 4514–4523.
- Xia, J., S. Kalluri, C. Micheyl, and E. Hafter. 2017. Continued search for better prediction of aided speech understanding in multi-talker environments. *The Journal of the Acoustical Society of America* 142 (4): 2386–2399.
- Xia, J., N. Nooraei, S. Kalluri, and B. Edwards. 2015. Spatial release of cognitive load measured in a dual-task paradigm in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America* 137: 1888–1898.
- Zurek, P.M. 1993. Binaural advantages and directional effects in speech intelligibility. In *Acoustical Factors Affecting Hearing Aid Performance*, ed. G. Studebaker, and I. Hochberg, 255–276. Needham Heights, MA: Allyn and Bacon.

Applying Cognitive Mechanisms to Audio Technology

Creating Auditory Illusions with Spatial-Audio Technologies



Rozenn Nicol

Abstract Perception of sound fields reproduced by loudspeaker arrays which are driven by spatial-audio technologies, such as Wave-Field Synthesis or Higher-Order Ambisonics, is examined in the light of “*Spatial Auditory Illusions*”. The spatial-audio technologies are based on illusions as to which real sound-sources vanish perceptually in favor of virtual sources. Furthermore, spatial-audio technologies are able to synthesize sound fields which are very similar to sound fields as created by real sources. In this chapter, these illusions are first explored as a function of the acoustic (physical) properties of the synthesized sound fields. Then, the perceptual dimensions are reviewed of what is actually heard when being exposed to these synthesized sound fields.

1 Introduction

What is meant by “*Spatial Audio*”? This term denotes any technology of sound reproduction based on loudspeaker arrays, in particular the following ones: *stereophony*, *quadraphony*, *multichannel audio* as, for example, defined by the surround-sound standards 5.1, 7.1, 9.1, 10.2 and 22.2 (Rumsey 2018), further 1st-order Ambisonic, Auro3D®, and sound-field synthesis (SFS) methods such as *wave-field synthesis (WFS)*, *higher-order Ambisonics (HOA)* and *near-field-compensated higher-order Ambisonics (NFC-HOA)*—and, more generally, all other methods of sound-field control (Spors et al. 2013; Zhang et al. 2017).

In all of these technologies, the goal is to create a sound scene with spatialized audio components. Usually they are separated into two categories, namely, on the one hand, technologies for achieving physical reconstruction of the sound field and, on the other hand, technologies that take advantage of psychoacoustic effects to alleviate the effort of the reconstruction of perceived sound fields. The latter technologies rely clearly on auditory illusions—in the sense that the listeners’ perceptions are

R. Nicol (✉)
Orange Labs, Lannion, France
e-mail: rozenn.nicol@orange.com

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_20

581

manipulated and misled. However, the former technologies are not free from illusions either—as will be discussed below.

The objective of this chapter is to go deeper into the understanding of the perception of sound fields as created by spatial-audio technologies. The concept of auditory illusions, which originates from studies of Auditory Scene Analysis in pure laboratory experiments (Deutsch 1983), will be revisited for the specific case of spatial-sound reproduction. In a first step, the concept of *Spatial Auditory Illusion* (SAI) is introduced.

2 Perception of Spatial-Sound Reproduction: More or Less Nothing But Illusions

2.1 *Sound Scene versus Auditory Scene*

At this point, it is important to distinguish the *sound scene* from the *auditory scene*. As defined by Blauert (1996), the former corresponds to the acoustical (physical) phenomena (i.e., the sound waves), whereas the latter is the perceptual interpretation of the sound events by the listeners. Stereophonic reproduction can be taken as an illustration here. Two loudspeakers create two sound events which are, however, generally perceived as one single auditory event—colloquially called “*phantom source*” in the field. This auditory event is localized at an intermediate position between the two loudspeakers. As pointed out by Linkwitz (2007), this is “a rather amazing phenomenon that has no precedence in the gradual evolution of natural hearing. For example, not sufficiently often have the sounds from two roaring lions been similar enough to locate as one lion somewhere between the two.” This observation holds for all the other technologies of spatial audio. Each loudspeaker generates a sound event which is generally not perceived as such. At the entrance of the listeners’ ears, the contributions of all loudspeakers are combined with various amplitude and delay relationships. All this information is processed by the listeners to elaborate auditory scenes. The extraction of single auditory events¹ from the eardrum signals is referred to as *Auditory Scene Analysis* (ASA) (Bregman 1990; Deutsch 1983). The psychoacoustic and cognitive mechanisms involved in this step are complex and remain not well understood.

Modeling of these processes is proposed by *Computational Auditory Scene Analysis* (CASA) (Wang and Brown 2006). Analysis of auditory scenes (Fig. 1) is a combination of both bottom-up (i.e., signal driven) and top-down (i.e., hypothesis driven) processes (Blauert 1999). Bottom-up algorithms perform well for localizing and tracking multiple sources if reverberation is low. As soon as reverberation increases, top-down processing is needed, namely, a-priori knowledge (i.e., rules

¹*Auditory events* must be distinguished from *auditory objects*. The latter result from a higher level of analysis and integrate potential cross-modal information (Wierstorf 2014).

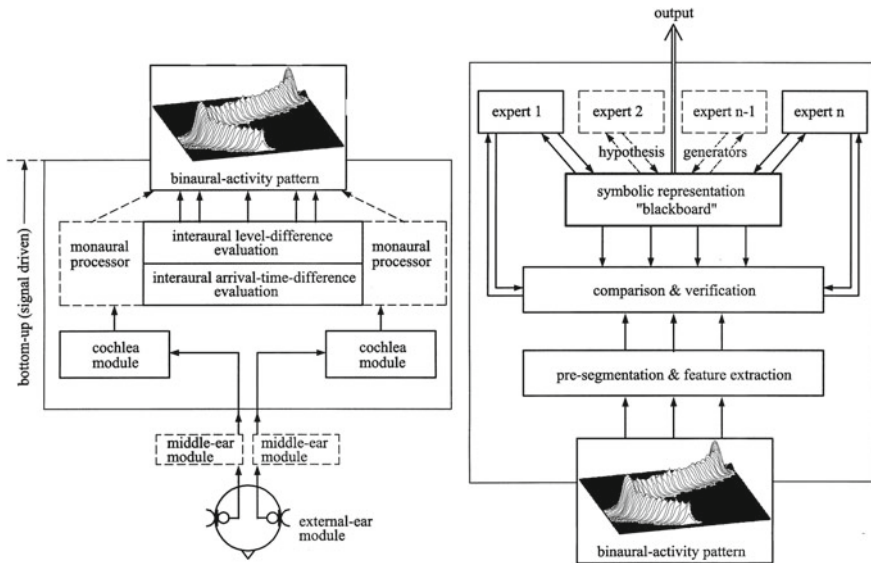


Fig. 1 Elaboration of the auditory scene from acoustic input. Reproduced courtesy of Jens Blauert (Blauert 1999)

or databases) is included from which hypotheses are generated. Each hypothesis is tested and finally accepted or rejected. As a result, an interpretation of the sound scene is provided.

2.2 Unconscious Inferences

Starting from the assumption that the physical model of the world represents the “real” world which “causes” the perceptual world, the following interpretation is commonly accepted. Humans do not sense physical objects directly but rather the signals that they emit (like light or sound). Thus, any percept in natural hearing or sound reproduction is a construct, that is, a specific representation of the physical reality, which is elaborated (or imagined) from the output of sensory receptors. As pointed out by Helmholtz in his “sign” theory of perception, sensations *symbolize* physical stimuli and are not a direct copy of them (Patton 2016). The correspondence between sensation and physical object is learned through experience. This process of interpretation was called “unconscious inferences” by Helmholtz. Consequently, the representation should not be mistaken for the physical world.

In natural perception, sensory illusions occur when this process of construction defaults. In this case, information from the physical world is misinterpreted,² which

²Illusions may also be caused by ambiguous sensory input.

leads to potentially strong discrepancies between the perceptual and the physical scene. The origin of this confusion relies on the process of “unconscious inferences” described by Helmholtz, and which can improve the efficiency of perception in most situations, but may fail in some cases as soon as the assumptions of inferences are not valid, as illustrated by auditory illusions, such as, for example, identified by Deutsch (Deutsch 1983; Warren 1983).

Moreover, the perceived result may depend on the specific listener, since the mechanisms involved in its construction partly rely on individual cognition. For instance, when investigating auditory illusions, Deutsch showed that the illusion may strongly vary from one listener to another (Deutsch 1983). Handedness, that is, individual preference for using the right or left hand is one important factor—potentially in relation to brain dominance.

2.3 *The Concept of “Spatial Auditory Illusions”*

To begin with, the following fact should be realized. *In spatial-sound reproduction the perceived sound scene is fundamentally and exclusively an illusion.* Indeed, a set of acoustic sources (i.e., loudspeakers) is used to create sound events, but it is clearly intended that, in the listeners’ minds, these physical sources vanish in favor of the auditory scene as expected by semantic or artistic intention. For example, in the case of stereophony, Lipshitz states that “One sign of a good recording/reproduction system [...] is that the loudspeakers are not audible as sources of the sound” (Lipshitz 1986). Instead of loudspeakers, the auditory scene is composed of virtual sources which have no physical support and are totally dissociated from the electroacoustic sources. Producing *Spatial Auditory Illusions* (SAI), that is, illusions of auditory scenes with audio components spatially distributed around the listener, is thus inherent to the spatial-audio technologies.

Not only does the localization of sources matter, but also the acoustic interaction with their environment, that is, with the acoustic space, including reflection, diffusion or diffraction phenomena in relation with the source directivity. The illusion involves both the creation of virtual sound-sources and spaces, and the manipulation of their perceived properties, for instance, positions and spatial extents of sources, attributes of room effects.

The following sections will investigate the properties of SAI, both acoustically and perceptually. Firstly, the acoustic signals are examined which are produced at the entrances to the listeners’ ears, and from which the illusion is generated. The objective is to better understand the relationships between the sound fields and the illusions. In other words, it is a question of identifying the features of the acoustic signals that can affect the illusions, and the type of auditory processes involved in the formation of the illusion. Secondly, leaving acoustic stimuli aside, SAI will be considered on their own, focusing on their perceptual properties. One fundamental issue in this context is the effectiveness of the illusions, in other words, the questions of to what extent do they function in an intended way, and what are the most effective technologies

for this purpose? As a prerequisite for observing and characterizing the perceptual properties of illusions, tools, and methods used for the perceptual assessment of spatial-audio technologies will be briefly described.

3 Illusion Versus Holophonic Reconstruction: Is Exact Reconstruction of the Sound Field Achievable?

As previously explained, an auditory illusion is a perceptual phenomenon that originates in the interpretation of the acoustic signals that are delivered to the entrance of the listeners' ears. Before going deeper into the perceptual mechanisms, the current section will focus on the acoustic properties of the sound fields as generated by spatial-audio technologies to the end of analyzing the links between the sound fields and the associated illusions. Intuitively, the most straightforward way to create effective illusions consists in reproducing the sound field exactly as it would be apparent in natural listening. This is avowedly the goal of SFS, such as aimed at in WFS or HOA installations. This idea is discussed in the following.

3.1 *Stereophony and Surround-Sound Systems*

In the early stage of spatial audio, two opposite strategies were investigated with the pioneering work of Blumlein and Fletcher (Lipshitz 1986). One, followed by Blumlein at E.M.I. (Electric and Musical Industries), was based on a pair of coincident microphones and a two-loudspeaker system, providing the basis of stereophony. The other, studied by Fletcher at Bell Telephone Laboratories, was using a “curtain of microphones” to record the sound sources. For the reproduction step, the microphones were connected to a “curtain of loudspeakers”. The superposition of the wavelets emitted by each loudspeaker restored the original wavefront, in a way close to that which will later be formalized as “holophony” by Jessel (Jessel and Vogel 1973). Fletcher’s aim was to re-create “the original macroscopic acoustic wavefront within the listening environment.”

With only two loudspeakers, this is clearly out of the reach of stereophony, the goal of which is rather to “re-create the wavefront on a microscopic scale”, that is, in a limited area around the listener’s head. This area is known as the so-called *sweet spot*. Because of these limitations, any design of stereophony needs to include sound perception to achieve its purpose. Besides, it can be stated that stereophony is fundamentally based on an illusion. It was referred to as the *auditory perspective* by Bell’s researchers. In other words, “the best that stereo can do is to provide a credible illusion that between and beyond the pair of loudspeakers there exists another acoustic environment within which the musicians are located and performing. [...] The only major question is how to produce at the listener’s ears from the two source

loudspeakers such differences as will be interpreted by the hearing mechanism as representing a credible image between the loudspeakers” (Lipshitz 1986).

Surround sound systems were motivated by the need to reproduce spatial audio for large audiences, such as in movie theaters, for which two loudspeakers were no longer sufficient (Rumsey 2018). Adding a central loudspeaker stabilizes the central image, whereas adding loudspeakers at the back and sides of the listener allows the system to reproduce sound ambiance, for example, isolated sound effects and reverberation. This can be seen as a first attempt to immerse the listener in a “sound field resulting from direct and reflected waves arriving from all directions”, and thus to get closer to “true realism” (Woodward 1977). However, in most surround systems, the loudspeaker layout is not regular, such as in the 5.1 surround setup. Consequently, not all directions are reproduced with equal accuracy. In general, frontal sources are privileged to the detriment of rear and side sources. One exception is quadraphony.

The localization of phantom sources is controlled by laws derived from two-channel stereophony, namely, by acting on interchannel differences of arrival time and level (Rumsey 2018). Finding the appropriate way to drive the loudspeakers was not straightforward, however. Pairwise panning is the simplest solution to generalize stereophony to a multi-loudspeaker configuration, but localization is altered as soon as the listener moves away from the central position. Furthermore, experiments have shown that the creation of phantom sources by pairwise panning works best between the left and right front loudspeakers. Performances are noticeably degraded for rear loudspeakers, and even more so for side loudspeakers (Cabot 1977). More optimal methods make use of all the loudspeakers, such as the “Cooper-and-Shiga” law for quadraphony (Woodward 1977). It quickly became clear that localization theories of stereophony cannot directly be applied to multiple-loudspeaker systems. The risk of creating a “confusion-phonetic system” instead of stereophonic sound (Willcocks and Badger 1983) was pointed out.

3.2 *Spatial Aliasing*

Today, holophonic reconstruction for spatial-audio reproduction is effective with technologies such as WFS, HOA, and general methods of sound field control. However, an exact reconstruction of wavefronts for the whole audible range is still illusory (Spors et al. 2013; Spors and Ahrens 2008; Wierstorf 2014). Spors and Ahrens (2008) proposed a unified formulation of WFS and HOA sound-field reproduction showing that, in both approaches, sound-field reconstruction theory is based on a continuous distribution of loudspeakers. In practice, discretized loudspeaker arrays are used which, however, causes spatial aliasing. Consequently, the sound reproduction is altered for frequencies higher than the aliasing frequency, f_{al} . This frequency depends on the loudspeaker spacing. For the typical spacing in the range of 10 to 20 cm, the aliasing frequency lies between 1 and 2 kHz (Ahrens and Wierstorf 2015). The consequences of spatial aliasing are different for WFS and HOA. A reason for this is that the spatial spectrum of the sound field is band-limited in the case of

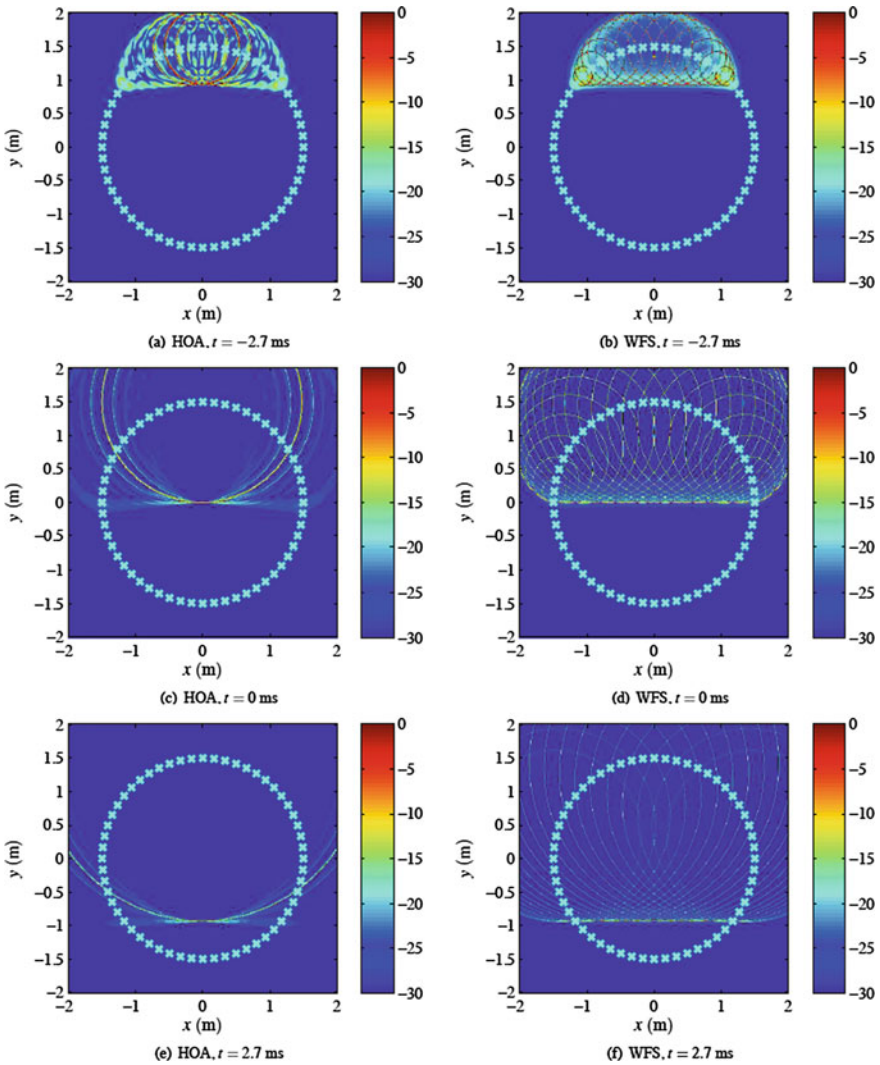


Fig. 2 Illustration of sound-field reconstruction by HOA (left column) and WFS (right column). Impulse responses of the loudspeaker system in the horizontal plane (amplitudes expressed in dB SPL). Reproduction of a plane wave propagating downwards by a circular array composed of 56 equi-angularly spaced loudspeakers. Reproduced courtesy of Jens Ahrens, with permission of the Audio Engineering Society, www.aes.org (Ahrens et al. 2010)

HOA (Spors and Ahrens 2008), thus preventing spatial aliasing. Nevertheless there remains a reconstruction error that is associated with spatial aliasing (Ahrens et al. 2010).

For instance, if the reproduction of a plane wave by WFS (anechoic environment) is examined, a first broadband wavefront is observed, the shape of which is accurately synthesized over the whole listening area, but which is initially followed by a dense sequence of broadband “wavelets” within a temporal range of 0–0.2 ms, and then by a slightly sparser sequence of high-frequencies wavelets within a range of 0.2–6 ms. These wavelets, resulting from spatial aliasing, come from various directions and spread over the entire area behind the wavefront—see Fig. 2. They originate from the individual contribution of each loudspeaker. Therefore their time distribution depends on the size of the array. For small arrays the last wavelet arrives a few ms after the first wavefront, for instance, 5 ms for a linear array of 3.1 m length, and 25 ms for a linear array of 12.7 m length (Ahrens and Wierstorf 2015).

In contrast, in the case of HOA, the first wavefront is accurately reproduced over a smaller area in the vicinity of the center of the listening area. It should be noticed that this wavefront conveys only the low-frequency components. It is followed by a short sequence of high-frequency wavelets, which are concentrated on its immediate neighborhood (Fig. 2). Thus, the major difference to WFS reproduction is that there is a temporal separation between the low- and high-frequency content³ which, moreover, are coming from different directions. The low-frequency wavefront propagates well in the desired direction, whereas the high-frequency wavelets arrive from the directions of the loudspeakers.

In the frequency domain, the consequence of spatial aliasing is that the transfer function of the system, as observed at a given listening point, is strongly altered above the aliasing frequency, exhibiting many ripples similar to comb-filtering (Ahrens and Wierstorf 2015). Furthermore, the spectrum distortions depend highly on the listening position. In addition, it should be remarked that, below the aliasing frequency, the frequency response is not flat due to approximations introduced by WFS. This deviation can be compensated by pre-filtering the signals that drive the loudspeakers, but the correction will only be effective over a restricted part of the listening area. Generally, near-field reconstruction is favored over far-field reproduction.

3.3 Properties of the Accurate Reconstruction Area

When observing the reproduction of a monochromatic plane wave as a function of frequency (Spors and Ahrens 2008), it is observed that the wave is accurately synthesized at low frequencies. As the frequency increases, artifacts similar to interferences occur for both WFS and HOA. The consequence is that the area of accurate reconstruction shrinks as a function of frequency for both WFS and HOA. However, as noted above, the properties of the reconstruction errors differ remarkably between the two technologies. In the case of HOA, the area of accurate reproduction remains in the center of the loudspeaker array, which is thus free of artifacts whatever the

³This separation is probably related to the spatial-bandwidth limitation.

frequency, whereas, in the case of WFS, it moves towards the direction opposite of the most active loudspeakers.

3.4 Additional Errors

Spatial sampling is not the sole deviation from the ideal holophonic reconstruction. Instead of planar arrays, linear arrays (i.e., 2.5D reproduction) are generally used. This leads to an incorrect amplitude decay as a function of distance (Sonke et al. 1998; Wierstorf 2014). Further, with linear arrays of loudspeakers, truncation effects occur since the size of the array is necessarily limited. The consequence is a reduction of the area of accurate reproduction and the creation of additional sources at the edges, resulting in diffracted wavelets (Spors and Ahrens 2009). In addition, it should be remarked that accurate reproduction of very low frequencies is not possible. To minimize the spacing between loudspeakers, small loudspeakers are preferred to the detriment of their low-frequency response. To overcome this limitation, one or several subwoofers may be used, but the sound field reproduction can still only be corrected at one specific position that is defined as the reference point (Ahrens and Wierstorf 2015). Finally, approximations that are introduced in practical implementations of WFS or HOA are another cause of deviations from an ideal reconstruction of the sound field. One example is the fact that it is assumed that loudspeakers are omnidirectional, which real sources are not.

3.5 Reproduction of Focused Sources

Synthesis of focused sources is a specific case of sound-field reproduction (Spors et al. 2009; Ahrens and Spors 2009; Ahrens and Wierstorf 2015). A focused source is a virtual source which is located downstream of the loudspeaker array. The sound field first converges to the source location (i.e., the focus point), and then diverges as if it were created by a source at this position. In the area between the loudspeakers and the source location, the reconstruction of the sound field is erroneous, in the sense that the reconstructed wave propagates in the opposite direction to the expected one, that is, from the focus point to the loudspeakers. On the contrary, the wavefront synthesized downstream of the focus point is well reconstructed, with the exception of spatial aliasing. However, the temporal properties of the artifacts differ from those observed for non-focused sources. Actually, they precede the wavefront and thus lead to pre-wavelets—as illustrated in Fig. 3. It should be noticed that the time distribution of these pre-wavelets strongly varies as a function of the lateral position (i.e., the x-coordinate) of the listener. Furthermore, the longer the array, the more pre-wavelets there are. It has therefore been suggested to decrease the length of the loudspeaker array to minimize the artifacts when synthesizing focused sources.

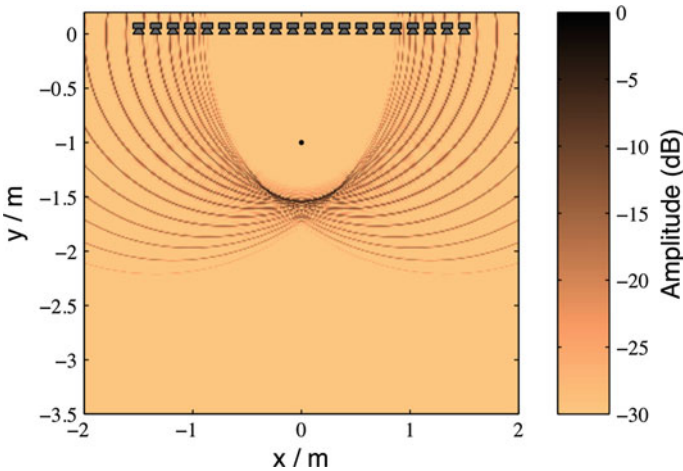


Fig. 3 Illustration of sound-field reproduction of focused sources by a linear array with a loudspeaker spacing of 10 cm. Impulse response of the loudspeaker system in the horizontal plane. Reproduced courtesy of Jens Ahrens, with permission of the Audio Engineering society, www.aes.org (Ahrens and Wierstorf 2015)

3.6 Room Reverberation

So far, the focus of the discussion has been put on the reproduction of direct sounds. The accuracy of the reproduction of room reverberation was not yet examined. In the case of HOA, there are two main ways to create HOA signals to drive the loudspeakers. One of them is to record natural sound scenes by 1st-order Ambisonics or HOA microphones (e.g., Soundfield® or Eigenmike® microphones). The other one is to derive synthetic HOA signals by processing pre-existing recordings of sound sources. In contrast to Ambisonics, WFS is exclusively based on virtual sound scenes represented by synthetic signals. In natural recordings, direct and reflected waves are naturally mixed, reverberation is therefore present. In synthetic signals, artificial reverberation is generally added, unless the reproduction space itself provides a perceptually satisfactory reverberation. Accurate reproduction of sound fields requires the introduction and fine control of artificial reverberation. The problem of creating early reflections is similar to synthesizing sound sources, that is, early reflections can be modeled as waves emitted by mirror images of the primary source (de Vries et al. 1994). All the limitations as were pointed out previously for non-focused sources apply for the reproduction of early reflections as well. In addition, potential interferences between early reflections and the wavelets resulting from spatial aliasing should be emphasized (Ahrens 2014; Ahrens and Wierstorf 2015). The fine time pattern of reflections is not only blurred by these wavelets (which follow both direct and reflected waves), but also preserving a pre-delay between the direct sound and early reflections is difficult.

Generally speaking, accurate reproduction of late reverberation has not yet been implemented. Rough modeling is preferred. One solution consists in synthesizing late reverberation by superimposing plane waves coming from directions equally distributed around the listener (Sonke 2000). A limitation of this method is that it does not allow the relative time of plane waves to be varied as a function of the listeners' positions. Yet, the method has been successfully applied for synthesizing room modes (Ahrens 2014).

Gari et al. (2016) present an interesting comparison between real sources (more precisely, an orchestra of loudspeakers) and their WFS reproduction as focused sources in a concert hall (Detmold Concert Hall). The physical properties of sound fields have been carefully examined both in terms of frequency response and spatial distribution as a function of time. Parameters of room acoustics have also been computed. It was concluded that for all these criteria, large deviations are observed between the target sound field and its WFS copy. In addition, listening tests showed that the differences are clearly perceptible.

3.7 Conclusion

It is now clear that spatial-audio technologies fail to perfectly reproduce the sound field of a given real-world scenario, at least in the current state of their development. Nevertheless, one may wonder whether this prevents them from creating SAIs. As stated above, perception is based on an interpretation of sensory input. Thus, as perception is able to resolve sensory ambiguities, it may as well overcome part of the errors in sound-field reconstruction.

4 Interaction Across Sensory Modalities

Sound-field reconstruction is not the only aspect that matters in natural hearing. Another one among others is, for instance, the listeners' ability to interact with the sound field, for instance via head movement or modification of the listening point. Moreover, real-life listening is generally a multi-sensory experience. However, most of the time, spatial-sound reproduction provides only auditory information, so that information from other modalities is potentially in conflict with the auditory scene. The fact, that the listener may be disturbed by this cross-modal mismatch, is confirmed by a study by Francombe et al. (2015) who investigated the attributes governing the perceptual differences between real and reproduced sound fields. Attributes related to the visual and tactile modalities are pointed out. Despite all this, the illusion may still be effective. This raises several questions.

A first issue is to what extent and in what way auditory illusions are affected by other sensory modalities, such as the visual, tactile, proprioceptive ones. Two cases should be distinguished. On the one hand, when information from other modalities is

congruent with the auditory scene, and on the other hand, when there is a mismatch between auditory stimuli and other modalities. For the case of mismatch, it was already shown that the perception of auditory stimuli can be influenced by visual stimuli, and vice-versa. A famous example is the “*ventriloquism effect*”, which occurs when an audio stimulus and a visual stimulus are presented at the same time, but spatially separated (Howard and Templeton 1966; Thurlow and Jack 1973). The perceived localization of the sound-source is modified, because the visual stimuli interfere with the acoustic ones. More precisely, the sound localization is captured by the visual modality. This illusion highly depends on the temporal and spatial disparities between the audio and visual stimuli (Slutsky and Recanzone 2001). In some rare cases, the sound may capture the vision, for instance, when the visual stimulus is blurred (Alais and Burr 2004).

Another example of cross-modal interaction between hearing and vision is the “*McGurk effect*”, by which the combination of a visual stimulus (lip movements corresponding to the syllable [ga]) and an audio stimulus (corresponding to the sound [ba]) results in an audiovisual percept which is neither [ga] nor [ba], but [da] (McGurk and Macdonald 1976). There are many other illustrations of such perceptual biases induced by acoustical or visual stimuli, such as the *audiovisual bounce-inducing effect* (Sekuler et al. 1997; Grassi and Casco 2010), and the *illusory flash* (Shams et al. 2000; McCormick and Mamassian 2008).

It should be noticed in the current context that hearing, as well as vision, can capture perception. Welch and Warren (1980) suggested that such capture is governed by the predominant modality associated to the task. For instance, for localization seeing is predominant with regard to hearing. Further, it should be borne in mind that situations with discrepancies between stimuli presented from the different sensory modalities do not necessarily lead to one single unified percept. In other words, the different sensory stimuli are only aggregated into one multi-sensory percept, if the conditions of perceptual fusion are met (Lewald and Guski 2003). These conditions are defined by a window-of-integration, the length of which depends specifically on the temporal and spatial disparities of the stimuli.

A further question concerns the possibility of auditory illusions to be sufficiently effective to infer an illusion in other modalities, that is, a kind of transfer of the illusion from one modality to another—compare, for example, Suzuki et al. (2020), this volume. All these issues deserve a deeper investigation to better comprehend the mechanisms of multi-sensory illusions.

5 Sound-Field Reconstruction in Light of Perception

When assessing the performances of a spatial-audio system, the observation of the reproduced sound field provides some information, but it is not sufficient to arrive at a comprehensive assessment (Spors et al. 2013). Perceptual evaluation is needed in addition, namely, not only to validate the physical results but also to provide new insights into the properties of spatial-audio reproduction. Furthermore, results from

physical and perceptual observation may be contradictory. Careful insight into these contradictions raises at least the following two questions.

- If exact reconstruction of sound fields were achievable, would it ensure perfect illusions, namely, an auditory scene free of artifacts and in full congruence with the expected result?
- Otherwise, would inaccurate reconstruction systematically lead to poor illusions?

These issues are now discussed in light of selected examples.

5.1 Is Perfect Reconstruction Desirable?

As has been explained above, exact reproduction of sound fields over both the whole audible bandwidth and over a wide listening area is still out of reach of current spatial-audio technologies. Perceptual assessment of perfectly reconstructed sound fields is therefore difficult. However, a study by Tucker et al. (2013) provides some surprising initial answers. A listening test was performed to evaluate the influence of spatial dithering (i.e., time misalignment of loudspeaker signals) on preference judgment of HOA reproduction (4th-order HOA reproduction by a 12-loudspeaker hemispherical array). Delays within a range of 0–8 ms or 0–15 ms were randomly applied to the set of loudspeakers. Four audio samples (busy restaurant, train station, public park, and church choir), recorded by an EigenMike® microphone, were assessed by pairwise comparison based on a preference judgment (forced choice). 24 participants (12 experts and 12 non-experts) took part in the experiment. The results suggest a tendency to prefer the dithered versions. By expert listeners, the misaligned version is preferred by around 73% of the 16 trials. Non-expert listeners do not exhibit a clear preference.

The paradoxical consequence is that perfectly accurate reconstruction of the sound field may not be a proper solution. However, this experiment is a pilot study, and the results deserve further investigation before drawing conclusions. It would have been interesting to ask the participants about the perceptual attributes that motivated their preference judgment. Moreover, one may wonder whether the authors of the study really succeeded in achieving a perfect temporal alignment of the loudspeakers. Small delays between the loudspeakers⁴ in the aligned version could then cause artifacts (e.g., coloration or phasiness), which are potentially more perceptible than in the misaligned version in which random delays tend to blur the artifacts. It has indeed been shown by Start (1997) that introducing random delays between loudspeakers can help to lower the perceptibility of spatial aliasing.

⁴Even though perfect temporal alignment of loudspeakers has been achieved, it is almost impossible to guarantee this alignment for the signals at the entrance of the listeners' ears, unless the listeners' heads are fixed.

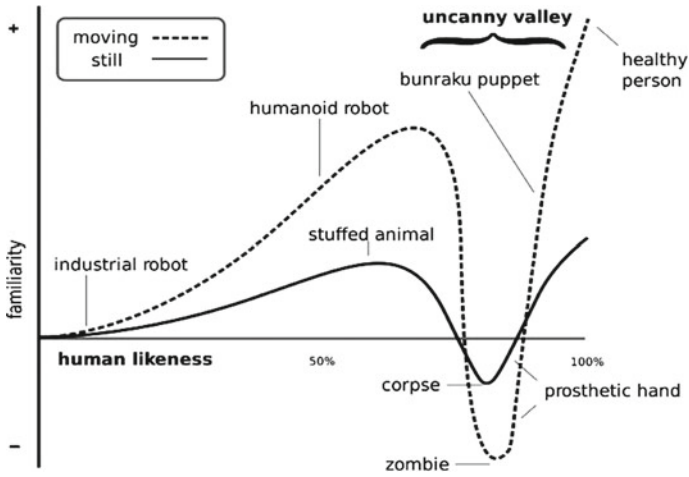


Fig. 4 Masahiro Mori’s “Uncanny Valley”. Reproduced courtesy of Glenn Dickins, with permission of the Audio Engineering Society, www.aes.org (Dickins et al. 2013)

5.2 The “Uncanny Valley” of Spatial-Audio Technologies

Instead of wondering whether perfectly reconstructed sound fields can reveal audible artifacts, a better question may be whether listeners become more sensitive to small deviations when sound field reproduction comes close to perfect reconstruction but without fully achieving it. Indeed, perceptual assessment of virtual-reality technologies suggests that people seem to be more tolerant of reproduction impairments when the quality is low or medium. On the contrary, when the rendering becomes very close to realism, the *Quality-of-Experience* (QoE) scores badly as soon as small defects are perceived. This phenomenon is referred to as the “*Uncanny Valley*”, namely, “the point where attempts to artificially reproduce human action and interaction come uncomfortably close to realism” (Dickins et al. 2013)—see Fig. 4. It should be remarked that this effect is potentially influenced by the rating paradigm for QoE. One may wonder whether direct comparison between the same impairments would lead to the same conclusions. Nevertheless, it is probable that the technologies of sound-field reconstruction, such as WFS or HOA, have reached the point of the uncanny valley.

In their article, Dickins et al. (2013) list up the factors which potentially cause an uncanny sensation in spatial-sound reproduction for interactive communication. (i), A first aspect is the rendering of space (*spatial fidelity*). Spatial distortion may be a “road into the uncanny valley”. (ii), Latency is another factor, namely when the congruence between close and far sources is lost, for instance, when the latter appear with too low a temporal lag. (iii), Time invariance contributes to the uncanny sensation as well. In natural experience, small movements of sources and listener cannot be avoided, leading to fine temporal variations of timbre or localization. It is

therefore recommended to add varying textures to overcome artificiality. (iv), The dynamic range, possibly reduced because of a low signal-to-noise ratio, may also have an impact. In particular, the lack of low-level signals (i.e., very fine sounds that are hardly audible) is a critical cue. (v), Eventually, perceptual continuity and situational congruence are mentioned. It should be highlighted in this context that the background components of the sound scene can be almost as important as the foreground ones.

5.3 *Illusions Despite Imperfect Reconstruction*

Another question of interest concerns poorly reconstructed sound fields, which may sound both pleasant and convincing in terms of spatial-audio imaging. Stereophony is exemplary for such situations. The sound field recreated by a stereophonic system over the whole listening area (i.e., the area encompassing the listeners' ears) departs noticeably from the one as induced by real sound sources. More exactly, contrary to binaural technology, in stereophony, it is not intended to reconstruct the acoustic signals at the entrance to the listeners' ears. Yet, the sound information delivered to the listeners is sufficient to convey the illusion of a sound scene in front of them (Leakey 1959; Lipshitz 1986).

Early work suggests that stereophony and, thus, the stereophonic illusion were rapidly accepted by the public (Leakey 1959). The principle of stereophony can be summarized as an appropriate manipulation of the relationships between the loudspeaker signals to the end of creating stereophonic illusions. As explained in Leakey (1959), Lipshitz (1986), amplitude differences between the loudspeaker signals produce *Interaural Arrival-Time Differences* (ITDs) between the left and right ears of the listeners in the sweet-spot area, particularly in the low-frequency range. Leakey observed that the interpretation of stereophonic scenes (i.e., the localization of the phantom sources) is mainly governed by ITDs.⁵ He also highlighted the importance of small movements of the listeners' heads, which provide variations of ITDs and allow the listeners to solve localization ambiguities.

Similarly, it has been shown in Sect. 3 that sound fields synthesized by WFS or HOA reveal inaccurate reconstruction of both time and frequency properties of the sound scenes (Spors and Ahrens 2008; Wierstorf et al. 2012; Spors et al. 2009; Geier et al. 2010; Ahrens and Wierstorf 2015), potentially leading to audible artifacts. Furthermore, the comparison of sound fields synthesized by WFS and HOA shows that the spatial and temporal properties of their aliased sound fields differ noticeably (Spors and Ahrens 2008).

In the case of WFS, wavelets resulting from spatial aliasing spread over the entire area behind the wavefront, whereas in HOA the aliased sound field is concentrated in the immediate vicinity of the main wavefront, suggesting that the effects of spatial

⁵Since ITDs are particularly effective in the low-frequency range, it is not a good idea to replace the two stereo loudspeakers by a common sub-woofer.

aliasing may be more perceptible for WFS than for HOA. However, in the case of HOA, there is both a temporal and a spectral separation between the low and high frequencies in reproduced sound fields. Perceptual merging of the low frequencies and high frequencies components is thus not guaranteed. Another potential consequence of spatial aliasing is that the reconstructed sound field may exhibit more or less variations as a function of the listening position, which may affect, for instance, the sound level or the spectral content of the virtual source. This leads to audible artifacts when the listener moves. It is therefore difficult to predict which technology will induce the most perceptible defects.

The question of the perceptibility of the wavelets and pre-wavelets resulting from spatial aliasing was examined by Ahrens and Wierstorf (2015) for the case of WFS. Although the wavelets are clearly detectable in visualizations of the reproduced sound fields, they are still likely to be inaudible. At least they are assumed not to affect the perceived location because of the phenomenon of *summing localization* (Blauert 1996), as long as the time delay between the main wavefront and the wavelets is smaller than about 1 ms. However, when the delay is in the range of 1–20 ms, the *precedence effect* (Blauert 1996) is acting, provided that the amplitude of the wavelets is low enough in comparison with the first wavefront. Under these conditions, wavelets are not perceptible. Yet, such components are probably associated with reflections and, consequently, contribute to room perception.

In the case of focused sources, pre-wavelets are observed. Contrary to wavelets, they are potentially highly perceptible. One reason is that they are unnatural. Another one is that they precede the main wavefront. Thus, their audibility is hardly decreased by the precedence effect and/or summing localization. Perceptual studies further show that strong colorations appear (Ahrens and Wierstorf 2015). In addition, the auditory event is split into two components, one localized at the focal point of the focused source and the other one (a high-passed version) at the nearest edge of the loudspeaker array. Click-like artifacts are also reported for long loudspeaker arrays. These artifacts depend on the nature of the signals, thereby signals with transients increase the perceived effects.

Surprisingly, localization accuracy is preserved (Wierstorf et al. 2012, 2017). The mean localization error measured for a WFS array with a loudspeaker spacing of 17 and 34 cm (corresponding to an aliasing frequency around 1 kHz and 500 Hz) is 1° and 3°, respectively, in the case of non-focused sources—Fig. 5. Moreover, localization accuracy remains good within the whole listening area delimited by the loudspeaker array. To explain these results, Ahrens and Wierstorf (2015) highlight the fact that a significant part of the information relevant for spatial perception is contained in the frequency range that is synthesized correctly.

NFC-HOA exhibits lower localization accuracy with mean errors of 3.8° and 7.4° for loudspeaker spacings of 17 and 34 cm, respectively. As for stereophony, estimation of interaural cues, that is, interaural arrival-time differences (ITDs) and interaural level differences (ILDs), brings further insight into these results (Fig. 6). Indeed WFS performs better in reconstructing a valid ITD for the low-frequency range (up to 1.3 kHz). Furthermore, in the case of 7th-order NFC-HOA reproduction,

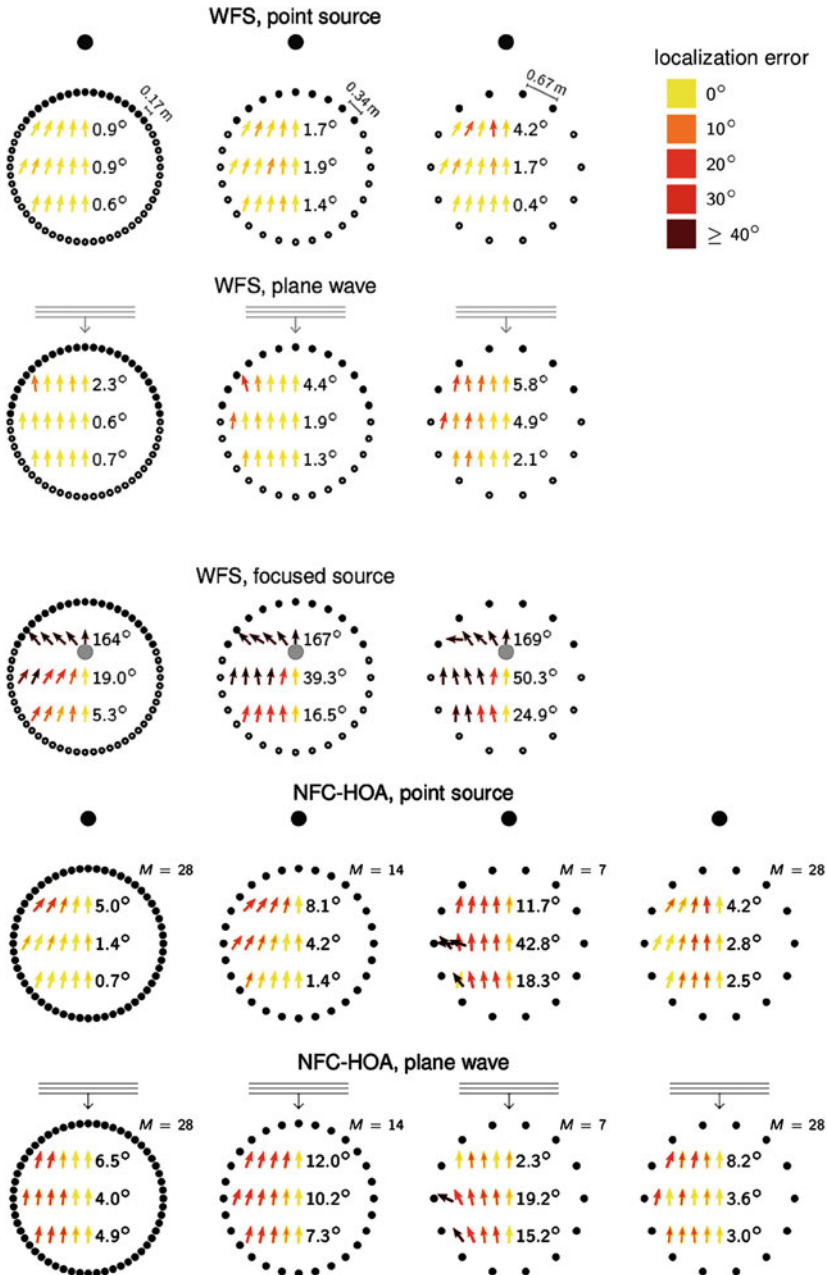


Fig. 5 Localization results measured for WFS and NFC-HOA sound field synthesis. The circular array of loudspeakers is depicted by **small circles**. The virtual source is depicted by a **large circle** (point source) or **three parallel lines** (plane wave). For each listening position, an arrow is pointing in the direction from which the listener perceives the auditory event. The color of the arrow displays the absolute localization error. The average error is indicated beside the arrows for every row of positions. **M** refers to the encoding order of HOA synthesis. Reproduced courtesy of Wierstorf et al. (2017)

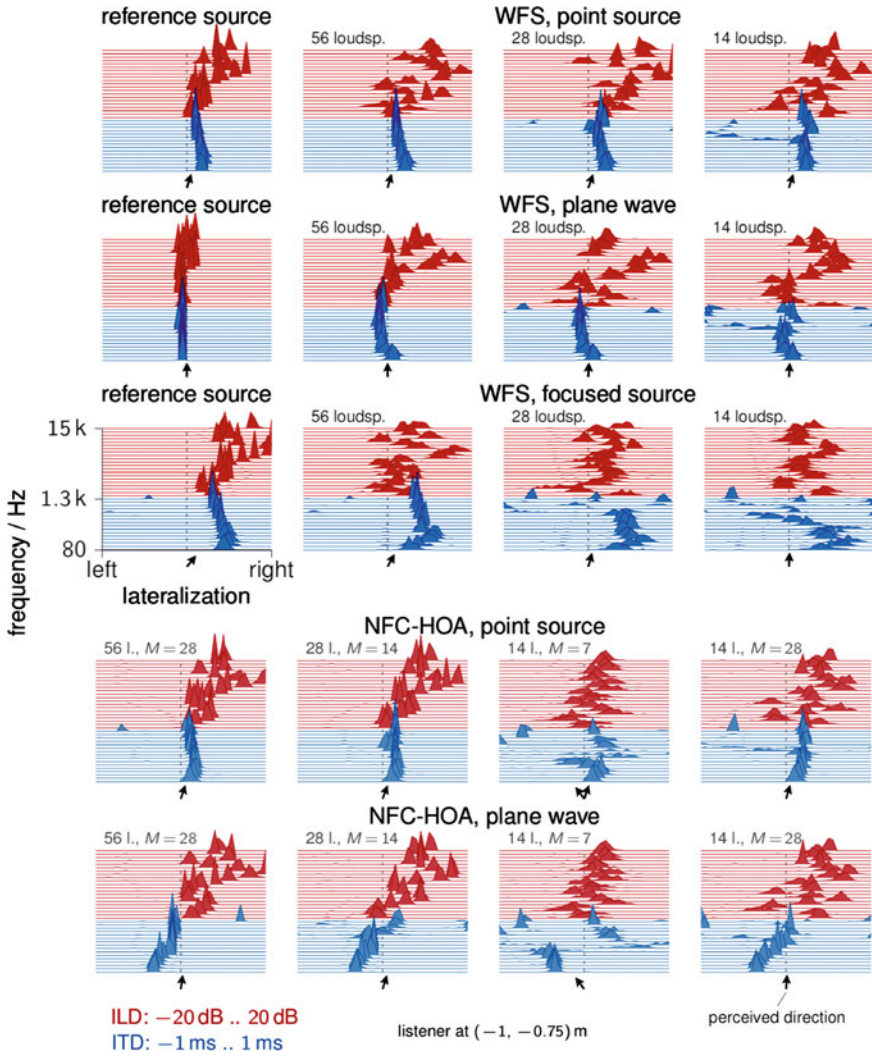


Fig. 6 ITD and ILD estimates for WFS and NFC-HOA sound field synthesis of one noise pulse. The ITDs were calculated up to 1.3 kHz, ILDs above 1.3 kHz. Each line displays the ITD or ILD histogram for one frequency band (vertical axis). Results are given for one listening position (-1 m, -0.75 m) in different sound fields. Reference source refers to the target sound field (point source, plane wave or focused source). ITDs and ILDs for these reference sources are compared with those obtained for the sound fields synthesized by WFS or NFC-HOA. Reproduced courtesy of Hagen Wierstorf (Wierstorf et al. 2017)

ITD values are spread out, causing a splitting-up of the virtual source into two components (Ahrens et al. 2010).

A recent study by Frank and Zotter (2017) revisits the definition of the sweet spot in the light of perception. The sweet-spot area can be physically defined as being delimited by a given threshold of the error of sound-field reconstruction. Frank and Zotter (2017), however, propose a new definition based on a perceptual criterion as follows. “The sweet-spot area is measured as the area within which the reproduced sound scene is perceived as *plausible*”. They designed an experiment in which the listener had to explore the listening area in order to identify the boundaries of the sweet-spot area. In the case of a dry audio scene, plausible means that localization of auditory events is preserved in front of the listener. In the case of a reverberant scene, plausibility judgment is based on both frontal localization of direct sound and envelopment of reverberation (Frank and Zotter 2017). For the experiment, a 2D-HOAI system composed of 24 loudspeakers was considered. 1st-, 2nd and 5th-order Ambisonics were compared. The result is that the sweet-spot area based on the plausibility criterion appears to be larger than the one based on the threshold of the physical error. As expected, its size increases with HOA order. The authors explain that HOA is more of an amplitude-panning method rather than a sound-field synthesis method. The consequence is that, as for stereophony, level differences between loudspeakers create interaural arrival-time differences at low frequencies. The localization of auditory events is determined by summing localization.

Further perceptual assessment of sound fields reproduced by spatial-audio technologies revealed another uncanny result. Garí et al. (2016) describe a listening test in which the perception of real loudspeakers and focused sources, synthesized by WFS in a concert hall, was compared. Preference judgments were collected. They show a significant tendency of preferring WFS, suggesting that illusions can be even better than reality. However, preference is a multi-factorial construct. A comprehensive evaluation including further perceptual attributes has to be performed before detailed conclusions can be drawn.

All these examples illustrate that the link between physical and perceptual properties of the sound field is not straightforward. Particularly, it appears that the various inaccuracies of sound-field reconstruction are not equivalent from a perceptual point of view. Actually, some of them are rated as strongly unpleasant, whereas others are not even noticed.

6 Spatial Audio in the Light of Auditory Scene Analysis

To better understand why there may be such a discrepancy between physical measurement and perception, the sound field created by a spatial-audio system must be examined more carefully. When a sound field is reproduced by a loudspeaker array, the acoustic pressure at any location is the result of the superposition of all the sound events emitted from all loudspeakers. Even when the resulting sound field is identical to a sound field produced by real sound sources, the situation is strongly

different in terms of auditory scene analysis (Bregman 1990). Furthermore, in natural hearing, most sound scenes are composed of many acoustical events, including reverberation. The auditory system has developed efficient processes to extract meaning⁶ from this complex situation. In other words, the auditory system has the ability to separate sound components or to reorganize them into streams. Consequently, our brain is potentially able to discriminate the contributions of individual auditory streams, which may lead to unexpected illusions. This is, for example, illustrated by the auditory illusions that Deutsch (1983) has reported.

6.1 *Summing Localization versus Association Model*

Theile suggested that such streaming phenomena also happen in stereophonic reproduction (Theile 1980, 1991; Wittek et al. 2007). He opposed the conventional theory of summing localization, as described by Blauert (1996), with a new model: the “*Association Model*”—(Fig. 7). The acoustic pressure observed at each ear of the listener is the sum of the two waves emitted by the left and right loudspeaker. The main idea of summing localization is that, instead of perceiving two sound events corresponding to the two loudspeakers, the listener perceives one single auditory event, i.e., the “phantom” source, which is localized as a function of the interaural relations of the resulting waves at the listener’s ears. The detailed computation of the time/phase and amplitude differences between the left- and right-ear signals as a function of the stereophonic signals driving the loudspeakers is, for example, given in Lipshitz (1986). These interaural differences are then interpreted in terms of localization cues, which leads to a direct relationship with the perceived localization of the phantom source (Blauert 1996). This theory relies on the assumption that the two sound components induced by the left and right loudspeakers at each ear are merged into a sole event and further analyzed as such, which is allowed by the fact that the signals of the two loudspeakers are coherent and differ only by small time and/or amplitude shifts.

This assumption is reconsidered by the association model. In this model, it is argued that the auditory system is able to separate the components of the left and right loudspeakers at each ear. The subsequent auditory processing is then decomposed into two steps. (i), The left and right ear signals are individually spatially decoded (i.e., inverse binaural filtered), thus allowing for extraction and localization of the contributions of the left and right loudspeakers—*location association*. (ii), The similarity of the loudspeaker signals leads to a merger that results in one single auditory event, namely, the phantom source. The latter process is based on high-level auditory scene analysis—*Gestalt association*. The time/phase and amplitude relations between the loudspeaker signals are extracted independently of the crosstalk between the loudspeakers and the ears, thanks to their spatial separation achieved during the first stage, the location association. One consequence is that the

⁶Meaning is what the auditory percepts mean to the listeners in their current situation.

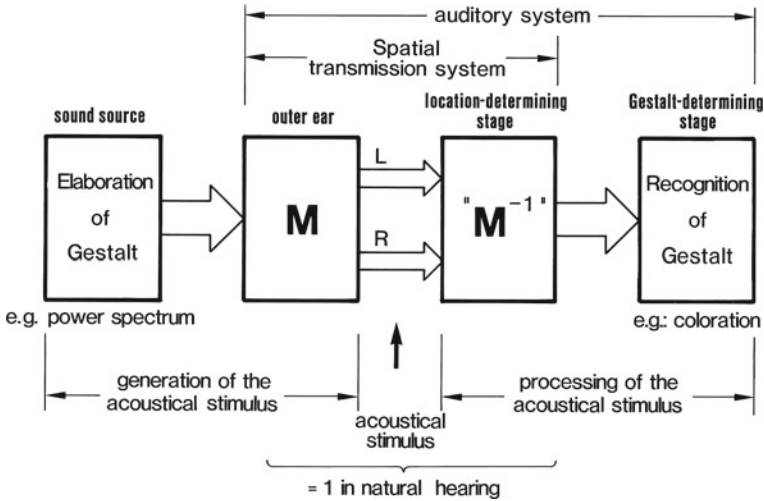


Fig. 7 Association model proposed by Theile. Reproduced courtesy of Günther Theile, with permission of the Audio Engineering Society, www.aes.org (Theile 1986)

perceived direction of the phantom source is no longer interpreted as a function of the interaural differences, but as a function of the relationships of primary stereophonic signals emitted by the loudspeakers. Another consequence is that, since the loudspeaker signals are properly separated and localized, comb-filter coloration, which would occur when summing up the delayed versions of the same signal at each ear, is suppressed (Wittek et al. 2007).

The association model suggests that the analysis of stereophonic signals combine both segregation and fusion of the elementary components of each loudspeaker. Segregation acts in the sense that the loudspeakers are identified as two sources localized at separate positions, thanks to the binaural differences associated with each source. Localization means that spatial decoding of the loudspeaker signals is performed, in that the *Head-Related Transfer Functions* (HRTFs) from the left and right loudspeaker to each ear are inverted. As a result, the differences introduced by the wave propagation between the loudspeakers and the listener’s ears are canceled, which increases the coherence between the two auditory streams and, consequently, leads to fusion. In other words, segregation is linked to spatial information, whereas fusion concerns the signal content itself—mainly spectral information. This model supports a distinction between an auditory “where” subsystem and an auditory “what” subsystem, as described in Kubovy and Van Valkenburg (2001). What is new in this theory is that the acoustic signal at one of the listener’s ear is not processed as a whole. Despite the fact that one single auditory event is finally perceived, an underlying separation is presumed to be effective during the process of perception.

Assuming that the association model is valid, the question arises of whether this observation also holds for multi-loudspeaker reproduction such as WFS or HOA. In

a more general sense, the question is what happens when the number of loudspeakers becomes greater than two? The issue of the auditory fusion of signals coming from multiple loudspeakers has already been raised by Willcocks and Badger (1983), when investigating the localization of phantom sources in early surround systems. This situation is similar but not equivalent to the perception of a sound source in a room, where plenty of sources are added in terms of reverberation. Binaural decoloration of the primary source is then observed despite the high number of coherent additional sources. However, as remarked by Wierstorf (2014), the association model cannot explain this phenomenon.

Indeed, Wierstorf observed that separation is also effective for WFS and HOA reproduction, suggesting that the auditory system is aware of the “superposed” nature of the acoustic input (i.e., the fact that it is composed of several elementary wavefronts), which is then interpreted accordingly. This result questions the validity of the binaural assessment of multi-loudspeaker reproduction of sound fields (Wierstorf et al. 2013). Indeed there may be a perceptual difference between the natural and the synthetic superposition of elementary waves. The method was validated in terms of localization, but the perception of other perceptual attributes was not yet assessed. A second question is whether there is a maximum number⁷ of loudspeakers beyond which separation is no longer possible? It is likely that, as the number of sources becomes too high, separation is made difficult, if not impossible.

This is a plausible explanation of why coloration or phasiness artifacts are reported for WFS reproduction, whereas coloration is judged less annoying in stereophony (Wittek et al. 2007). The aforementioned study of Tucker et al. (2013), which suggests a preference for dithered versions of HOA sound reproduction, supports the idea that the interpretation of sound fields reproduced by spatial-audio technologies would need increased separation—which is provided by intentional time misalignment in this case—between the individual contributions of the loudspeakers. Yet, a better understanding of the underlying mechanisms is required. Particularly, a thorough investigation of the factors (e.g., the number of loudspeakers, their interspacing and signal relations) that govern the separation or merging of loudspeaker contributions is necessary. Further, the influence of the current tasks and/or the cognitive states of the listeners should be evaluated (Wierstorf 2014).

6.2 *Spatial Re-Organisation of Auditory Streams*

The previous section provided a first insight into auditory scene analysis. It was realized that the interpretation of sound scenes may involve segregation or fusion of audio components. An important consequence is that these components are then spatially re-organized into streams, meaning that the perceived direction of one given component may be co-determined by other features than the location of the sound

⁷From speech recognition, it is known that humans can segregate up to 6 speakers in multi-speaker scenarios, i.e., in cocktail-party situations (May et al. 2013).

event, for example, by the spectral content. This is illustrated by the auditory illusions as studied by Deutsch (1983).

Auditory illusions are revealed in experiments on the perception of complex sound scenes. A complex sound scene is composed of at least two streams. Contrary to experiments based on isolated sound stimuli, perception of complex scenes shows the existence of powerful high-level mechanisms that may strongly modify the auditory percept in a way that has little to do with the sound event. Complex stimuli are more representative of natural hearing. In real-life listening, the interpretation of a sound scene may be ambiguous. To solve these ambiguities, additional information (i.e., assumptions, or top-down information) are employed. Moreover, in real-life situations, sound alterations are naturally interpreted as information, since these alterations reflect events related to the sound wave propagation, such as movement of sources, presence of obstacles. In Deutsch’s experiments, the sound stimuli were artificially modified, leading to the distortion of the associated percept. This distortion provides some insight into the mechanisms of auditory scene analysis. Some examples are discussed in the following.

Deutsch (1983) studied the perception of two streams, emanating from two different locations. In practice, the experiments were based on headphone listening. This allowed precise control of the signals perceived by the left and right ears. The “Octave Illusion” was observed in presence of two sinusoid signals, one at 400 Hz and the other one at 800 Hz. These signals were presented alternately to the left and the right ear—Fig. 8a. Most listeners perceived two streams, namely, a sequence of higher tones on the right ear, alternating with a sequence of the lower tones on the left ear—see Fig. 8b. This percept is a clear distortion of the sound stimulus.

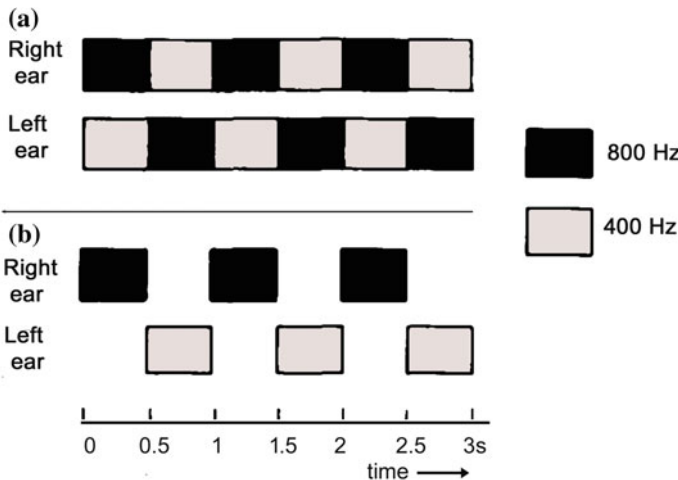


Fig. 8 The octave illusion—(a), sound stimuli presented to the listeners’ ears, (b), most commonly observed associated percept. Adapted courtesy of Diana Deutsch, with permission of the Audio Engineering society, www.aes.org (Deutsch 1983)

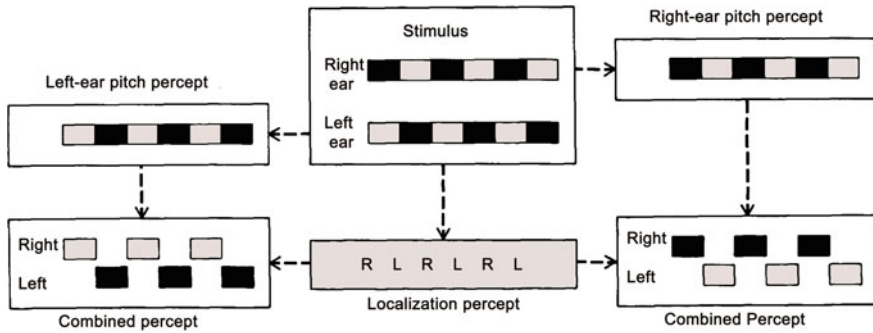


Fig. 9 The octave illusion—Model describing the two mechanisms involved, one of them determining the pitch and the other one the perceived location. Adapted courtesy of Diana Deutsch, with permission of the Audio Engineering Society, www.aes.org (Deutsch 1983)

The illusion depended on the individual listener. Alternative perceived patterns were reported. Individual differences have been found to correlate with handedness, that is, complex patterns were most often observed among left-handers. Indeed, these results can be partly explained by left/right hemisphere dominance. Even more troubling is that the illusion was generally not modified if the left and right signals were reversed, confirming that the perceived location depended very little if at all on the location of the sound events. Furthermore, the octave illusion was also observed despite crosstalk, that is, with the sound stimuli being presented over two loudspeakers, separated or arranged side-by-side in an anechoic environment. When the listeners turned their heads, the illusion was maintained, except when facing one loudspeaker. An analog illusion was achieved with complex tones. However, reverberant environments reduced the effect. Parameters potentially influencing the percept, such as frequency interval, time delay, and relative amplitude, were carefully examined by Deutsch.

The model proposed by Deutsch explains the octave illusion as the combination of two separate mechanisms (Fig. 9). One of them is responsible for pitch determination (i.e., to determine “*what*” pitch is heard). The other one takes care of sound localization (i.e., to determine “*where*” the auditory event is positioned in the perceptual space—compare (Kubovy and Van Valkenburg 2001)). The perceived pitch results from a mechanism of side dominance, by which the listener focuses on the information in one ear, for instance, the right one. Any information presented to the other ear is then suppressed. The mechanism responsible for localization is governed by frequency, that is, the auditory event appears in the ear that perceives the highest frequency. Further investigation suggested that the ear dominance effect is useful to counteract interferences by reflections and reverberation, as in the case of the precedence effect. As regards the frequency effect on sound localization, it may be due to acoustic shadowing by the listeners’ heads, which low-pass filters the signals perceived by the contralateral ear. Therefore, in natural listening, the ear which receives the highest frequencies is naturally associated with the source direction. New

evidence of these two separate processing steps for the identification of the sound stimuli and for the spatial location of the auditory events has been contributed by observations by means of *functional Magnetic Resonance Imaging* (fMRI)—Zatorre et al. (2002), Wang and Kuriki (2012).

A second example of spatial re-organization is the *Scale Illusion*. A sequence of ascending frequency tones is presented to the listener, with successive tones alternating from one ear to the other. A second sequence of descending frequency tones is presented simultaneously in the same way but to the opposite ear. The most often reported pattern consists of two separate melodies, one of them composed of the highest tones and localized in the right ear, and the other one composed of the lowest tones and localized in the left ear. In this illusion, localization is exclusively governed by frequency, resulting in merging if the frequency is similar, and in segregation otherwise. The usefulness of this mechanism is justified to prevent room reflections from blurring localization. First-order cues of localization are thus discarded in favor of secondary cues, such as content similarity. Eventually, Deutsch (1983) studied the influence of the temporal relationships between the signals presented to the left and the right ear. One example is the situation where the melody is accompanied by a drone (i.e., low-frequency musical accompaniment). She showed that the ability to identify the melody, meaning that the listener separates the melody from the drone (i.e., the “*drone effect*”) depends on the temporal relationships between the left and right signal. This illustrates that temporal separation helps to segregate sound components. Indeed, in natural listening, sources that are spatially separated create strong temporal separation in the signals perceived by the listener.

The main lesson that can be learned from these observations is that, in certain situations, primary cues of localization can be overwritten due to high-level processes of auditory scene analysis (segregation or fusion), leading to a perceptual remapping of sound components (Moore 2007). As illustrated by the octave and the scale illusions, the similarity of the frequency content has been shown to act as a very powerful cue for grouping, despite the spatial separation of the acoustic sources. These auditory illusions have several potential implications for spatial-audio technologies. Reproduction of primary cues of sound localization appears less essential than at first sight. For instance, reproduction inaccuracies of spatial information can be overcome by perceptual-grouping laws (e.g., spectral similarity). A better knowledge of auditory scene analysis could, therefore, help to improve spatial sound reproduction. The major limitation is the dependency on the individual listener, which is more or less unavoidable whenever perception is involved.

The predominance of spectral content over spatial cues questions the relevance of studies that examine the benefit of spatial-audio reproduction, such as WFS or HOA, to provide or enhance source unmasking, either with or without the visual modality being involved (Sanson 2011; Palacino et al. 2016; Vilkaitis and Wiggins 2017). Spatial parameters must be assessed in comparison with the spectral content and other parameters that may govern the auditory scene analysis. The laws governing the fusion or the segregation of loudspeaker signals as a function of their inter-relationships (e.g., similarities in the spectral contents and arrival-time or amplitude

differences) should also be investigated for a better understanding of the perception of spatial position and timbre.

7 Perceptual Assessment of Spatial-Audio Technologies

In the previous section, auditory illusions were examined on the basis of the sound field present at the entrance of the listeners' ears, more precisely in the light of the various processes (both bottom-up and top-down) in which the percept is elaborated from acoustic inputs. Now the focus of the discussion will be put on the percept by investigating the perceptual properties of the auditory scenes generated by spatial-audio technologies. But, to do this properly, tools are needed to observe and describe SAIs. A straightforward way is perceptual assessment of the results of spatial-audio reproduction. This section will present an inventory of the various methods and tools that are dedicated to measure the perceived features of reproduced spatial-sound scenes. Further details can be found in Raake and Wierstorf (2020), this volume.

7.1 *Experimental Parameters*

A preliminary remark is necessary regarding the experimental conditions of perceptual assessment. When reviewing studies on the perception of spatial-sound reproduction, it is seen that the material under assessment is strongly heterogeneous (Rumsey 2002). Often a single sound source is considered rather than a complex scene with many audio components. The audio content can be realistic (i.e., representative of signals in everyday life) or of laboratory style (such as random noise or sinusoidal signals). The reproduction system can also be real or simulated. For instance, multi-channel audio systems based on complex reproduction setups, e.g. with an excessive number of loudspeakers, are sometimes emulated by binaural synthesis (preferably dynamic binaural synthesis, i.e., in combination with head-tracking) (Wittek et al. 2007; Wierstorf et al. 2013). Most of the time, pure audio stimuli are used, but the influence of other sensory modalities is also be considered in some cases. All these parameters noticeably influence the character of the perceived auditory scene.

7.2 *Perceptual Attributes*

The perception of spatial-audio reproduction is multidimensional, which means that it is governed by many attributes. Identifying these attributes is a key question, investigated by many studies for more than 40 years. Although a wide and sometimes confusing variety of attributes has been proposed, some general trends can be observed. At least, attributes can be grouped into a limited number of categories,

which are quite similar from one study to another. Berg and Rumsey (2003) point out three main perceptual dimensions: (i), timbral attributes (relating to the tone color), (ii), spatial attributes (relating to the three-dimensional nature of the sound sources and their environments) and (iii), technical attributes (relating to distortion, hiss, hum, etc.). However, it has been suggested that timbral attributes dominate the rating of basic audio quality over spatial attributes by 70 over 30% (Rumsey et al. 2005).

In an attempt of standardization, the *Spatial-Audio-Quality Inventory* (SAQI) proposes a consensus list of 48 attributes for the perceptual assessment of virtual sound environments (Lindau et al. 2014; Lindau 2014). The verbal descriptors were gathered by a focus group of 20 German-speaking experts. The vocabulary was then also translated into English and French. One year later, in Zacharov and Pedersen (2015), 401 attributes were collected from 22 studies. A method of semantic text mining was applied to identify common meanings among attributes. These were then classified into five main clusters, namely, (i), spatial (distance and depth, presence, spatial impression, clarity, reverberance, width), (ii), timbre (coloration, tone color, sharpness, hardness, warmth), (iii), loudness, (iv), artifacts, and (v), hedonic (appealingness, naturalness, pleasantness, beautiful/ugly, subjective preference, degree of liking, realism).

Alternatively, two main categories can be distinguished, namely, (i), attributes that are related to a physical property of either the sound source, the acoustic space or the sound reproduction system (e.g., timbre, source location, source width, room effect), and (ii), affective attributes that concern the way in which the listeners' psychological or emotional state is modified by sounds (Nicol et al. 2014).

All these perceptual attributes are used in psychoacoustic experiments. Listeners are asked to rate a selection of audio samples with a subset of attributes. Ratings based on these attributes lead to both a qualitative and a quantitative representation of auditory scenes. Perceptual attributes thus provide an evaluation grid allowing to describe the perceived features of SAIs.

7.3 Models

Instead of running perceptual-assessment experiments for collecting listeners' judgments, models can be used to predict attribute ratings from acoustic signals. The input are either the signals that drive the loudspeakers, or the acoustic pressure measured at the entrance to the listeners' ears. For spatial-audio reproduction, models focus on spatial attributes, in most reported experiments the spatial positions of virtual sound sources. One of the early models was proposed by Makita (1962) for the estimation of the perceived direction of the phantom source created by a stereophonic system. In his general meta-theory of localization, Gerzon (1992) extended Makita's model to the case of multichannel sound reproduction. He introduced two criteria, namely, the velocity vector and the energy vector. These can be interpreted as spatial barycenters of incoming acoustic waves at the listening position. The vector direction points

to the perceived direction of the virtual source. Its magnitude measures the spatial spread of the reproduced sound field, namely, if its norm is close to one, it means that the sound energy is focused on a few loudspeakers.

The velocity vector is computed from the amplitude and phase signals of loudspeakers and is therefore similar to a low-frequency criterium, whereas the energy vector can be seen as its high-frequency version, in which the energy of loudspeaker signals is considered instead and the phase relationships are discarded. The velocity and energy vectors are commonly used to assess HOA reproduction. These localization criteria were reformulated by Kearney (2013) for off-center listening positions and used to assess the perceived height achieved by HOA sound-field reproduction (Kearney and Doyle 2015).

Binaural auditory models have also been developed to reproduce bottom-up and top-down processing of the auditory system (Kohlrausch et al. 2013; Søndergaard and Majdak 2013). One category of models focuses on the prediction of perceived localization. For instance, the binaural-activity map (Takanen et al. 2014a) represents the instantaneous activation as a function of time and frequency on a left-right 1D-map. Takanen et al. (2014b) showed that binaural-activity maps can be exploited for an estimation of the localization error at various listening positions (both center and off-center) in WFS and NFC-HOA reproduction. Moreover, binaural-activity maps allow visualizing coloration artifacts. Further modeling has been proposed in the TWO!EARS project (<http://twoears.eu>) for both localization and coloration prediction (Raake et al. 2014; Raake and Wierstorf 2016).

7.4 *Indirect Assessment*

Rating perceptual attributes is a direct assessment, that is, the listeners are fully aware that they are performing an evaluation of an auditory scene. The alternative is indirect assessment, in which no evaluation tasks are formally assigned to the participant. Information about the perceived features of the auditory scene is inferred by observing reactions and behavior of the test listeners (Faure 2005; Gonot 2008; Guillon 2009). For instance, the listeners can be asked to perform a given task in relation to the sound scene (e.g., to count the sound components, or to memorize some information). Their success rate is then interpreted in terms of the auditory-scene features. In such experiments, it has, for instance, been shown that the response time to localize virtual sources in binaural synthesis is a measure of the quality of the HRTF set (Guillon 2009).

Furthermore, neurophysiological measurements (e.g., electroencephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy (fNIRS), and Peripheral Autonomic Nervous System (PANS) signal acquisition) have been used to obtain insight into brain activity or physiological parameters such as heart rate, blood pressure, eye movement, skin conductance, respiration features. EEG/MEG is based on measuring the electrical/magnetic activity along a scalp, whereas fMRI and fNRIS observe

blood flows that accompany neuronal activity. EEG and MEG provide data with a high time resolution in the millisecond range, but with poor spatial resolution. By contrast, high spatial but low time resolutions are achieved by fMRI and fNIRS. All these methods have a high potential for investigating the perception of reproduced sound fields. It should be remarked that fMRI measurement is not compatible with multi-loudspeaker reproduction because of its operating noise and the risk of magnetic interferences with the loudspeakers, but this limitation can be overcome with sound-field reproduction simulated by binaural synthesis (Wiestorf et al. 2013), in combination with appropriate headphones.

Neurophysiological methods are considered as a promising way of investigating auditory perception, for instance, in the field of quality-of-experience (QoE) assessment (Akhtar and Falk 2017; Laghari et al. 2013). Brain activity was observed with EEG during an evaluation of speech degradation (Antons et al. 2010). Their results provided evidence that noise that is not perceived on a conscious level (i.e., behavioral data do not indicate that the stimulus is perceived as degraded) is nevertheless processed subconsciously in a certain percentage of the trials. This suggests that EEG can be used to detect minimal differences in audio assessment. A study by Gupta et al. (2016) showed that some EEG features are correlated with emotion primitives (i.e., valence and arousal). They were successfully used to predict the influence of human factors (i.e., users' perception, emotional and mental state) on a QoE score for a comparison of text-to-speech and natural speech. In a similar study, preference judgments were related to fNIRS features. More specifically, activation of OFC (orbitofrontal cortex) signals was observed in a case of valuation-based decision making (Laghari et al. 2014).

However, so far there have only been a few studies that implemented these methods to assess spatial-audio technologies. A first study compared spatial-response fields of the primary auditory cortex to virtual sound sources synthesized with individual and non-individual HRTFs in ferrets (Mrsic-Flogel et al. 2001). It was shown that the responses obtained with an animal's own ears differed significantly in shape and position from those obtained with another ferret's morphology. More recent studies have confirmed a positive correlation between various levels of accuracy of spatial sound reproduction (e.g., individual HRTF vs. generic HRTF vs. impoverished localization cues (Palomäki et al. 2005; Wisniewski et al. 2016), natural versus artificial cues of auditory motion (Getzmann and Lewald 2010), individual binaural versus stereo recordings for sound externalization (Callan et al. 2013)) and the activity of the auditory cortex measured either by MEG, by EEG or by fMRI. Most of the results reported above were obtained by using binaural synthesis. Sound-field reproduction by multi-loudspeaker systems has not yet been investigated in this context.

Future studies of this kind might be used in two ways, that is, either to observe the activity of the auditory cortex or to measure the affective state of the listener (i.e., emotion, preference). Beyond the comparison of neurophysiological responses with auditory illusions and real sound sources, all these methods are of particular interest for a deeper understanding of the perceptual mechanisms of auditory illusions.

8 Perceptual Properties of Spatial-Audio Illusions

In the following, the properties of auditory percepts created by spatial-audio reproduction are investigated by analyzing the results of perceptual studies. Firstly, those perceptual attributes are considered that have been shown to govern our perception of auditory scenes rendered by spatial-audio technologies—see Sect. 7.2. Many dimensions are involved here, such as the sound sources on their own (identity, spectral content and semantics of the signal radiated, etc.), the spatial properties of these sources (location, movement, directivity, size, etc.), and the acoustic environment as conveyed by reflected sound (size and type of the room, presence of absorbing, reflecting or diffracting surfaces, etc.).

It is now discussed how these perceptual attributes are rendered and controlled in SAIs, with a particular focus on spatial-related attributes. However, an additional question arises at this point of the discussion, that is, to what extent is the listener fooled? Controlling the localization of virtual sources is one issue, but another issue is to check whether the listener is aware of the “artificiality” of the illusion. Since auditory illusion means that sensory data are manipulated to infer a given percept that is more or less distinct from the physical reality,⁸ it is necessary to evaluate to what extent the listener is aware of this manipulation. In other words, one should ask whether the illusion is successful. Some perceptual attributes are correlated to this question, for instance, the *authenticity* or the *plausibility* of the auditory scene, also its *naturalness*, and its *presence* (Lindau et al. 2014). This issue is examined in a second step.

8.1 Spatial and Timbral Attributes of Spatial-Audio Illusions

The method of *Repertory-Grid Technique* (RGT) has been used to identify perceptual attributes relevant for multichannel sound (Berg and Rumsey 1999). From 6–17 attributes were collected for each individual listener. The predominant categories concern artifacts (e.g., clean sound vs. chirpy, squeaky, unnatural sound) and localization. Other attributes were associated with coloration (original vs. filtered, balanced vs. unbalanced frequency response), distance (far vs. close) and reverberance (dry vs. reverberant). *Principal-components analysis* (PCA) of attribute ratings suggested that the perceptual space is governed by two main dimensions, one of them being related to a combination of artifacts and distance, the other one related to source localization. Perceptual dimensions of WFS reproduction were also explored in the particular case of focused sources (Geier et al. 2010; Wierstorf et al. 2013).

⁸A kind of “trompe-l’oeil” that is, cheating, yet, here not of the eyes but of the ears.

Azimuth

Evaluations of spatial-audio reproduction often concern the source localization, that is, most studies focus on the perceived azimuth of virtual sources. The accuracy of azimuth localization has been extensively investigated in sound fields reproduced by both WFS and NFC-HOA (linear and circular arrays)—Wierstorf et al. (2017, 2012). The localization error was measured for virtual point sources, plane waves, and focused sources. The influence of the listening position was also examined. WFS reproduction achieves perceptibly lower errors than HOA. The mean errors are less than 2° for a loudspeaker spacing of around 20 cm and 3° for a 40 cm spacing. This is close to the localization accuracy observed for real sound sources. The localization accuracy in WFS remains good within the whole listening area, but it decreases significantly for focused sources, where the mean error is generally greater than 10° . For most of the conditions, NFC-HOA reproduction exhibits larger errors than WFS. That is, in NFC-HOA the mean error is 3.8° for a 17 cm loudspeaker spacing and 7.4° for a 34 cm spacing). Furthermore, localization accuracy depends on the listener position.

In addition, as it is known that higher-order Ambisonics components improve the accuracy of sound-field reproduction, one may then wonder to what extent the accuracy of localization is also enhanced. Several studies have examined this question (Braun and Frank 2011; Bertet et al. 2013), comparing 1st-order to 4th-order HOA systems, and showing a significant increase of localization accuracy as a function of the maximum order of Ambisonics components. However, the order of the highest components that are required to ensure an accuracy equivalent to the localization of real sources is still an open issue.

Elevation

Localization in the vertical plane was also assessed for WFS (de Bruijn 2004; Rohr et al. 2013) and HOA (Pieleanu 2004). The height of virtual sources is effectively reproduced by both systems.

Distance

Although earlier work questions the ability of WFS to reproduce sound distance in the specific case of focused sources (Wittek et al. 2004), more recent studies assessing the perceived distance of sound sources synthesized by WFS (Moulin et al. 2013a, b; Rébillat et al. 2011, 2012) and HOA (Kearney et al. 2012) have shown that distance perception is effective in both technologies. Lopez et al. (2014) compared distance rendering of WFS and *Vector-Base Amplitude Panning* (VBAP) and confirmed that WFS is more efficient than VBAP for the simulation of a sense of distance. As with azimuth, localization performances for both height and distance are close to what is achieved with real sources.

Width

Besides, it has been shown that the perceived width of a virtual source created by a 2nd-order HOA system can be controlled by appropriate filtering, namely, by introducing wave dispersion (Zotter et al. 2014). In the case of sound fields synthesized by WFS, Nowak et al. (2013) assessed the perception of two attributes, namely, ASW (*Apparent Source Width*), that is the spatial extent of the auditory event, and LEV (*Listener Envelopment*). These attributes are related to the perceived quality of rooms and, consequently, that of spatial sound fields. It was observed that their perception in virtual sound fields is similar to that in real sound fields as produced in concert halls, in particular, regarding the influence of early and late reflected sounds.

Timbre

In addition to the spatial properties of auditory scenes, timbre is another attribute receiving a lot of attention in perceptual assessments of spatial-audio technologies. Indeed, since the sound field reproduced results from the delayed superposition of several elementary waves associated to the different loudspeakers, comb-filtering is likely to occur and may lead to timbre distortion (i.e., coloration). This applies to stereophony as well to WFS or HOA, but the perceived effect depends on the number of loudspeakers. As illustrated by Wittek et al. (2007), who compared the perception of coloration in stereophonic and WFS reproduction, timbre distortion is stronger for WFS than for stereophony. A solution was proposed by the authors with the *Optimized Phantom Source Imaging* method, in which high frequencies (i.e., above the aliasing frequency) are reproduced by a subset of the WFS array. This allows for a reduction of the perceived coloration. Nevertheless, as noticed in Wierstorf et al. (2014), the perception of coloration is relative, that is, coloration is always evaluated in comparison with a reference. If no direct comparison is available, it may not be perceptible and the timbre is judged plausible.

In the same way, coloration is more audible if the listener or the source moves. Wierstorf et al. (2014) measured the perceived coloration (rating of perceived timbral differences based on a MUSHRA paradigm) of a WFS system as a function of the loudspeaker spacing within the range of 0.3–67 cm, and of the listening position. The experiment revealed that coloration was highly perceptible for most of the experimental conditions. For a noise stimulus, coloration was always audible, even for the smallest spacing, and it increased dramatically as soon as the loudspeaker spacing exceeded 4 cm. For speech stimuli, coloration only disappeared when the inter-loudspeaker distance was as low as 0.3 cm, but became strongly perceptible when the spacing was larger than 17 cm. The perceived coloration was lowest in the center of the listening area and remained homogeneous elsewhere.

8.2 *Effectiveness of Illusion into Question(s)*

The primary goal of spatial-sound reproduction is to create or recreate a sound scene, but the actual goal is to provide the listeners with the illusion that they are immersed in a sound scene. Among all the perceptual attributes that govern the perception of SAIs, one dimension that is particularly relevant for the illusion is whether and to what extent SAIs are effective. The term *effective* is used here in the sense that the listeners are fooled in such a way that they truly believe that the auditory scene is an exact representation of the physical reality which is presented to them. This dimension can be described in terms of “credibility”, or “effectiveness”, or “success” of the SAI.

An associated issue concerns the rating scale of this dimension. Is it a binary one—“yes”, if the SAI is successful, or “no”, if it is not—or a polar, gradual scale between the judgements “poor illusion” and “strong illusion”? This problem is probably not so simple as it looks at a first glance since the effectiveness of SAIs cannot be summarized in one single dimension. As Francombe et al. (2015) have stated, the perceptual differences between real and reproduced sound fields are clearly multi-dimensional. Responses of experienced listeners reveal more than 20 categories of perceptual attributes. Surprising as this may seem, this issue has so far been poorly addressed. The main reason is methodological because it is difficult to implement a direct comparison between natural and synthetic sound fields. Nevertheless, the results of the aforementioned study open promising insight into this issue. First attempts to connect the question of illusion effectiveness to existing attributes are proposed in the following.

8.3 *The Illusion of Completeness*

It should be realized that sound fields as created by WFS or HOA are highly heterogeneous. As has already been noticed above, they result from the superposition of many more or less different elementary waves. Therefore, the primary challenge of SAI is to create an illusion of completeness (Martens and Woszczyk 2007). For instance, despite the fact that the sound field is reproduced by a discontinuous array of loudspeakers, the auditory scene has to be perceived as continuous, which is achieved most of the time. But notice, for example, that in the case of stereophony, the sound scene does not fill all the space but is limited to the area between the two loudspeakers and behind them.

8.4 *Authenticity versus Plausibility versus Familiarity*

Another issue to be questioned is whether an illusion must conform to a reference to be rated as successful. In other words, is it required that the auditory scene matches a given scene to ensure the effectiveness of the illusion? For instance, in the case of the reproduction of an existing scene, the illusion is considered fully effective if the listeners' experience is identical to their experience in the corresponding real situation. To measure the effectiveness of the illusion, a straightforward method is to compare the reproduced scene with the reference as defined by the real scene. It turns out that a crucial concern in this context is the question of what denotes the term "*authenticity*" of the reproduced sound scene (Spors et al. 2013). If no explicit reference is available, the *plausibility* of the reproduced scene will be assessed instead (Spors et al. 2013). Such a reference-free assessment relies on implicit references, which depend on related past experiences of the listeners, and may also be affected by their expectations in general. Even for the same individual listener, the judgment can change over time because expectations may change. In other words, these (internal) references are not easily controllable, if at all. Thus, the judgment of plausibility as a substitute of authenticity may come out as strongly individual.

To repeat the above arguments in different wording, natural listening may provide an explicit reference (i.e., the "illusion of reality"), but implicit references can be based both on real or reproduced sound fields, which means that reality is not the sole reference. Specifically, spatial-sound reproduction (e.g., stereophonic reproduction), which include noticeable discrepancies from reality, may be used by the listeners as their references. As suggested by Rumsey (2002), the fundamental property of the references governing our judgment of plausibility (e.g., naturalness) is that there is "something" that the listeners have heard before.

As a consequence, unfamiliar sounds may be judged as "non-plausible", even though they were taken from real scenes. For instance, since most of our listening experience stems from reverberant scenes, an anechoic sound field may be perceived as unnatural. Multi-loudspeaker systems that are able to surround the listener with sounds coming from various different directions at the same time, which is rarely experienced in real life, create an auditory situation that potentially leads to unfamiliarity, and thus to uneasiness. Plausibility is affected whenever something is missing in the auditory scene, such as room effects, or something is in excess of everyday listening—"hyper-reality", such as a sound space saturated in all directions.

8.5 *"You Are There" versus "They Are There"*

In addition to the question of authenticity or plausibility, the attribute *presence* is relevant when investigating the effectiveness of SAIs. Rumsey (2001) distinguishes two categories of auditory illusions, that is, (i), the illusion of "You are there", meaning that the listeners feel as if they are in the place where the sound scene was

recorded (e.g., in a specific concert hall) and, (ii), the illusion of “They are there”, in which case they have the feeling that the sound sources are in their individual acoustic space, for instance, in their living room. From the point of view of sound engineering, this distinction is only relevant as an aesthetic choice. Yet, in terms of illusion effectiveness, this distinction clearly matters. The illusion of “You are there” implies that the listeners consciously and actively feel transferred into an illusionary scene that is in total contradiction to their knowledge and the information from sensory modalities other than hearing. This illusion requires that they accept to be virtually moved to another place. On the contrary, the illusion of “They are there” does not require any conscious cooperation of the listener. At any time, unexpected sound events may occur in their personal environment, which makes the occurrence of new sound sources plausible without extra cognitive effort.⁹ For this reason, the illusion of “They are there” is stronger than the illusion of “You are there”. Consequently, the former is more difficult to achieve. Early reflections and reverberation, as well as the interaction between the source directivity and the environment, need to be properly reproduced to create the illusion that virtual sources share the same acoustic space as the listener.

9 Conclusion

In this chapter, SAIs have been examined both in terms of acoustic wave reconstruction and perception. Although an exact copy of natural sound fields is still out of reach of even the most accurate technology of spatial audio, successful illusions may be achieved. Indeed natural perception is also a reconstruction, as even if the acoustic information is degraded, the auditory scene may sound convincing. A critical issue remains the lack of cross-modal information in pure audio reproduction. Further assessment of SAIs is also needed for a better comprehension of the overall effectiveness of illusions, which has received little attention so far in comparison to what is already known about spatial or timbral attributes. Future progress in the understanding of how the brain works, and, particularly, of the complex processes leading to the construction of the auditory percept by the brain, will certainly provide new insights into this issue.

Besides, it is definitely worthwhile to consider whether perfect auditory illusions are always desirable. For virtual-reality applications, an exact copy of the sound scene is usually expected, for instance for learning (e.g., in a flight-, driving- or sports-simulator) or for gaming purposes. However, spatial-audio technologies are also used for artistic purposes (e.g., music, cinema, entertainment). In these contexts, a reasonable level of accuracy is often sufficient rather than authenticity to create the intended illusions. Thus, what has to be kept in mind is the following. Artistic

⁹However with the input from other sensory modalities (e.g., visual cues), the plausibility of these new sources may collapse. It can, therefore, be assumed that the listeners focus exclusively on auditory information, for instance by closing their eyes.

work is motivated by the desire to create sensations and to convey emotions into the observers' minds. So above all, please note that

What actually matters is the illusion and not the accuracy of reproduction

Acknowledgements The author would like to thank Francis Rumsey for his very enlightening conversations that have strongly inspired this chapter. The author wishes to thank I. Viaud-Delmon for her thorough proofreading of the very first version of the manuscript and her encouraging feedback. The author wishes also to thank L. Gros, O. Warusfel, P. Rueff, L. Simon, P. Guillon, S. Moulin, J. Faure, J. Moreira, G. Roussel, M. Paquier, V. Koehl and J. Palacino for their invaluable collaboration in studies which have contributed to this chapter. Further thanks go to the two anonymous reviewers for their constructive comments, which greatly improved the readability of this chapter. This chapter is dedicated to the memory of my father, Yves Nicol (1928–2019).

References

- Ahrens, J. 2014. Challenges in the creation of artificial reverberation for sound field synthesis: Early reflections and room modes. In *EAA Joint Symposium on Auralization and Ambisonics*, Berlin, Germany.
- Ahrens, J., and S. Spors. 2009. Spatial encoding and decoding of focused virtual sound sources. In *Ambisonics Symposium*, Graz, Austria.
- Ahrens, J., S. Spors, and H. Wierstorf. 2010. Comparison of higher-order ambisonics and wave-field synthesis with respect to spatial discretization artifacts in time domain. In *40th International Conference: Spatial Audio: Sense the Sound of Space*. Tokyo, Japan: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=15563>.
- Ahrens, J., and H. Wierstorf. 2015. Properties of large-scale sound field synthesis. In *57th International Conference: The Future of Audio Entertainment Technology - Cinema, Television and the Internet*. Hollywood, CA, USA: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=17609>.
- Akhtar, Z., and T.H. Falk. 2017. Audio-visual multimedia quality assessment: A comprehensive survey. *IEEE Access* 5: 21090–21117. <https://doi.org/10.1109/ACCESS.2017.2750918>.
- Alais, D., and D. Burr. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* 14 (3): 257–262. <https://doi.org/10.1016/j.cub.2004.01.029>.
- Antons, J.-N., B. Blankertz, G. Curio, S. Møller, A.K. Porbadnigk, and R. Schleicher. 2010. Subjective listening tests and neural correlates of speech degradation in case of signal-correlated noise. In *129th Convention*. San Francisco, CA, USA: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=15628>.
- Berg, J., and F. Rumsey. 1999. Identification of perceived spatial attributes of recordings by repertory grid technique and other methods. In *106th Convention*. Munich, Germany: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=8256>.
- Berg, J., and F. Rumsey. (2003). Systematic evaluation of perceived spatial quality. In *24th International Conference: Multichannel Audio, The New Reality*. Banff, Canada: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=12272>.
- Bertet, S., J. Daniel, E. Parizet, and O. Warusfel. 2013. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta Acustica united with Acustica* 99 (4): 642–657. <https://doi.org/10.3813/AAA.918643>.
- Blauert, J. 1996. *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press.
- Blauert, J. 1999. Models of binaural hearing: Architectural considerations. In *Proceedings of the 18th Danavox Symposium*, Ballerup, Denmark, 189–206.

- Braun, S., and M. Frank. 2011. Localization of 3D ambisonic recordings and ambisonic virtual sources. In *ICSA*, Detmold, Germany.
- Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press.
- Cabot, R.C. 1977. Sound localization in 2 and 4 channel systems: A comparison of phantom image prediction equations and experimental data. In *58th Convention*. New York, NY, USA: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=3059>.
- Callan, A., D.E. Callan, and H. Ando. 2013. Neural correlates of sound externalization. *NeuroImage* 66: 22–27. <https://doi.org/10.1016/j.neuroimage.2012.10.057>.
- de Bruijn, W. 2004. Application of wave field synthesis in videoconferencing. Ph.D. thesis, Delft University of Technology (T.U. Delft), The Netherlands.
- de Vries, D., A.J. Reijnen, and M.A. Schonewille. 1994. The wave field synthesis concept applied to generation of reflections and reverberation. In *96th Convention*. Amsterdam, The Netherlands: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=6419>.
- Deutsch, D. 1983. Auditory illusions, handedness, and the spatial environment. *Journal of the Audio Engineering Society* 31 (9): 606–620. <http://www.aes.org/e-lib/browse.cfm?elib=4558>.
- Dickins, G., X. Sun, R. Cartwright, and D. Gunawan. 2013. The uncanny valley of spatial voice. In *52nd International Conference: Sound Field Control - Engineering and Perception*. Guilford, UK: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=16903>.
- Faure, J. 2005. Evaluation of dynamic binaural synthesis. Technical Report (Orange).
- Francombe, J., T. Brookes, and R. Mason. 2015. Elicitation of the differences between real and reproduced audio. In *138th Convention*. Warsaw, Poland: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=17731>.
- Frank, M., and F. Zotter. 2017. Exploring the perceptual sweet area in ambisonics. In *142nd Convention*. Berlin, Germany: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=18604>.
- Garí, S.V.A., J. Pätynen, and T. Lokki. 2016. Physical and perceptual comparison of real and focused sound sources in a concert hall. *Journal of the Audio Engineering Society* 64 (12): 1014–1025. <http://www.aes.org/e-lib/browse.cfm?elib=18535>.
- Geier, M., H. Wierstorf, J. Ahrens, I. Wechsung, A. Raake, and S. Spors. 2010. Perceptual evaluation of focused sources in wave field synthesis. In *128th Convention*. London, UK: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=15366>.
- Gerzon, M.A. 1992. General metatheory of auditory localisation. In *92nd Convention*. Vienna, Austria: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=6827>.
- Getzmann, S., and J. Lewald. 2010. Effects of natural versus artificial spatial cues on electrophysiological correlates of auditory motion. *Hearing Research* 259 (1): 44–54. <https://doi.org/10.1016/j.heares.2009.09.021>.
- Gonot, A. 2008. *Design and evaluation of interfaces to navigate into 3D sound environments (Conception et évaluation d'interfaces de navigation dans les environnements sonores 3D)*. Ph.D. thesis, Conservatoire National des Arts et Métiers, CNAM, Paris, France.
- Grassi, M., and C. Casco. 2010. Audiovisual bounce-inducing effect: When sound congruence affects grouping in vision. *Attention, Perception, and Psychophysics* 72 (2): 378–386. <https://doi.org/10.3758/APP.72.2.378>.
- Guillon, P. 2009. *Individualization of spectral cues for binaural synthesis: Looking for inter-individual similarities to adapt or reconstruct HRTF (Individualisation des indices spectraux pour la synthèse binaurale: recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF)*. Ph.D. thesis, Université du Maine, Le Mans, France.
- Gupta, R., K. Laghari, H. Banville, and T.H. Falk. 2016. Using affective brain-computer interfaces to characterize human influential factors for speech quality-of-experience perception modelling. *Human-centric Computing and Information Sciences* 6 (1). <https://doi.org/10.1186/s13673-016-0062-5>.
- Howard, I., and W. Templeton. 1966. *Human Spatial Orientation*. New York, USA: Wiley.

- Jessel, M., and T. Vogel. 1973. *Acoustique théorique. 1, Propagation et holophonie*. Masson, Paris, France.
- Kearney, G. 2013. Sound field rendering for distributed audiences. In *52nd International Conference: Sound Field Control - Engineering and Perception*. Guilford, UK: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=16902>.
- Kearney, G., and T. Doyle. 2015. Height perception in ambisonic based binaural decoding. In *139th Convention*. New York, NY, USA: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=17979>.
- Kearney, G., M. Gorzel, H. Rice, and F. Boland. 2012. Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields. *Acta Acustica united with Acustica* 98 (1): 61–71. <https://doi.org/10.3813/AAA.918492>.
- Kohlrausch, A., J. Braasch, D. Kolossa, and J. Blauert. 2013. An introduction to binaural processing. In *The Technology of Binaural Listening*, ed. J. Blauert. Springer and ASA Press.
- Kubovy, M., and D. Van Valkenburg. 2001. Auditory and visual objects. *Cognition* 80 (1–2): 97–126.
- Laghari, K., R. Gupta, S. Arndt, J.N. Antons, S. Møllery, and T.H. Falk. 2014. Characterization of human emotions and preferences for text-to-speech systems using multimodal neuroimaging methods. In *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, Toronto, ON, Canada, 1–5. <https://doi.org/10.1109/CCECE.2014.6901142>.
- Laghari, K., and R., Gupta, S. Arndt, and J.N. Antons, R. Schleicher, S. Møller, and T.H. Falk. 2013. Neurophysiological experimental facility for Quality of Experience (QoE) assessment. In *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, Ghent, Belgium, 1300–1305.
- Leakey, D.M. 1959. Some measurements on the effects of interchannel intensity and time differences in two channel sound systems. *The Journal of the Acoustical Society of America* 31 (7): 977–986. <https://doi.org/10.1121/1.1907824>.
- Lewald, J., and R. Guski. 2003. Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research* 16 (3): 468–478. [https://doi.org/10.1016/S0926-6410\(03\)00074-0](https://doi.org/10.1016/S0926-6410(03)00074-0).
- Lindau, A. 2014. *Spatial audio quality inventory (SAQI). Test Manual*. TU Berlin: Audio Communication Group. <https://depositonce.tu-berlin.de/handle/11303/157>.
- Lindau, A., V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl. 2014. A spatial audio quality inventory (SAQI). *Acta Acustica united with Acustica* 100 (5): 84–994. <https://doi.org/10.3813/AAA.918778>.
- Linkwitz, S. 2007. Room reflections misunderstood? In *123th Convention*. New York, NY, USA: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=14220>.
- Lipshitz, S.P. 1986. Stereo microphone techniques: are the purists wrong? *Journal of the Audio Engineering Society* 34 (9): 716–744. <http://www.aes.org/e-lib/browse.cfm?elib=5246>.
- Lopez, J.J., P. Gutierrez, M. Cobos, E. Aguilera. 2014. Sound distance perception comparison between wave field synthesis and vector base amplitude panning. In *2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, Athens, Greece, 165–168: <https://doi.org/10.1109/ISCCSP.2014.6877841>.
- Makita, Y. 1962. On the directional localization of sound in the stereophonic sound field. *E.B.U. Review* 73: 102–108.
- Martens, W.L., and W. Woszczyk. 2007. Illusions of music in space and spaces for music. In *UK 22nd Conference: Illusions in Sound*. Cambridge, UK: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=17291>.
- May, T., S. van de Par, and A. Kohlrausch. 2013. Binaural localization and detection of speakers in complex acoustic scenes. In *The Technology of Binaural Listening*, ed. J. Blauert. Chap. 15. Springer and ASA Press.
- McCormick, D., and P. Mamassian. 2008. What does the illusory-flash look like? *Vision Research* 48 (1): 63–69. <https://doi.org/10.1016/j.visres.2007.10.010>.
- McGurk, H., and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264 (5588): 746–748. <https://doi.org/10.1038/264746a0>.

- Moore, B.C.J. 2007. Perceptual organization of mixtures of sounds from different sources. In *UK 22nd Conference: Illusions in Sound*. Cambridge, UK: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=17276>.
- Moulin, S., R. Nicol, and L. Gros, 2013a. Auditory distance perception in real and virtual environments. In *Proceedings of the ACM Symposium on Applied Perception, SAP '13*, 117. New York, NY, USA: ACM. <https://doi.org/10.1145/2492494.2501876>.
- Moulin, S., R. Nicol, L. Gros, and P. Mamassian. 2013b. Subjective evaluation of audio egocentric distance in real and virtual environments using wave field synthesis. In *Proceedings of the Institute of Acoustics: Reproduced Sound, Manchester, UK*, vol. 35, 47–54.
- Mrsic-Flogel, T.D., A.J. King, R.L. Jenison, and J.W. Schnupp. 2001. Listening through different ears alters spatial response fields in ferret primary auditory cortex. *Journal of Neurophysiology* 86 (2): 1043–1046. <https://doi.org/10.1152/jn.2001.86.2.1043>.
- Nicol, R., L. Gros, C. Colomes, M. Noisternig, O. Warusfel, H. Bahu, B.F. Katz, and L.S. Simon. 2014. A roadmap for assessing the quality of experience of 3D audio binaural rendering. In *EAA Joint Symposium on Auralization and Ambisonics*, Berlin, Germany, 100–106.
- Nowak, J., J. Liebetrau, T. Sporer. 2013. On the perception of apparent source width and listener envelopment in wave field synthesis. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt am Wörthersee, Austria, 82–87. <https://doi.org/10.1109/QoMEX.2013.6603215>.
- Palacino, J., M. Paquier, V. Koehl, F. Changenet, and E. Corteel. 2016. Assessment of the impact of spatial audiovisual coherence on source unmasking. In *140th Convention*. Paris, France: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=18215>.
- Palomäki, K.J., H. Tiitinen, V. Mäkinen, P.J.C. May, and P. Alku. 2005. Spatial processing in human auditory cortex: the effects of 3D, ITD, and ILD stimulation techniques. *Brain Research. Cognitive Brain Research* 24 (3): 364–379. <https://doi.org/10.1016/j.cogbrainres.2005.02.013>.
- Patton, L. 2016. The Stanford Encyclopedia of Philosophy *Hermann von Helmholtz*, winter 2016 ed. Berlin: Springer. <https://plato.stanford.edu/archives/win2016/entries/hermann-helmholtz/>.
- Pieleanu, I.N. 2004. Localization performance with low-order ambisonics auralization, Master thesis, Rensselaer Polytechnic Institute, Troy, NY, USA.
- Raake, A., and H. Wierstorf. 2020. Binaural sound quality and quality-of-experience. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 393–434. Cham, Switzerland: Springer and ASA Press.
- Raake, A., and H. Wierstorf. 2016. Assessment of audio quality and experience using binaural-hearing models. In *Proceedings of the 22nd International Congress on Acoustics*. Buenos Aires, Argentina: ICA.
- Raake, A., H. Wierstorf, and J. Blauert. 2014. A case for TWO!EARS in audio quality assessment. In *Forum Acusticum*. Kraków, Poland: European Acoustics Association.
- Rébillat, M., X. Boutillon, E. Corteel, and B. Katz. 2011. Audio, visual, and audio-visual egocentric distance perception in virtual environments. In *Forum Acusticum*, Aalborg, Denmark, 482. <https://hal.archives-ouvertes.fr/hal-00619317>.
- Rébillat, M., X. Boutillon, E. Corteel, and B. Katz. 2012. Audio, visual, and audio-visual egocentric distance perception by moving participants in virtual environments. *ACM Transactions on Applied Perception* 9 (4): 19 (1–17). <https://doi.org/10.1145/2355598.2355602>.
- Rohr, L., E. Corteel, K.-V. Nguyen, and H. Lissek. 2013. Vertical localization performance in a practical 3-D WFS formulation. *Journal of the Audio Engineering Society* 61 (12): 1001–1014. <http://www.aes.org/e-lib/browse.cfm?elib=17077>.
- Rumsey, F. 2001. *Spatial Audio*. Massachusetts: Focal Press.
- Rumsey, F. 2002. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society* 50 (9): 651–666. <http://www.aes.org/e-lib/browse.cfm?elib=11067>.
- Rumsey, F. 2018. Surround sound. In *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*, ed. A. Roginska and P. Geluso. Routledge: New York, NY, USA. Chap. 6.

- Rumsey, F., S. Zielinski, R. Kassier, and S. Bech. 2005. On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *Journal of the Acoustical Society of America* 118 (2): 968–976. <https://doi.org/10.1121/1.1945368>.
- Sanson, J. 2011. *Musical and perceptual control of sound spatialization over an extensive area (Contrôle musical et perceptif de la spatialisation sonore en zone étendue)*, Ph.D. thesis, Université Paris Decartes, Paris, France.
- Sekuler, R., A.B. Sekuler, and R. Lau. 1997. Sound alters visual motion perception. *Nature* 385 (6614): 308. <https://doi.org/10.1038/385308a0>.
- Shams, L., Y. Kamitani, and S. Shimojo. 2000. Illusions: What you see is what you hear. *Nature* 408 (6814): 788. <https://doi.org/10.1038/35048669>.
- Slutsky, D., and G. Recanzone. 2001. Temporal and spatial dependency of the ventriloquism effect. *Neuroreport* 12: 7–10.
- Søndergaard, P., and P. Majdak. 2013. The auditory modelling toolbox. In *The Technology of Binaural Listening*, ed. J. Blauert. Chap. 2. Springer and ASA Press.
- Sonke, J.-J. 2000. *Variable acoustics by wave field synthesis*. Ph.D. thesis, Delft University of Technology (DUT), The Netherlands.
- Sonke, J.-J., J. Labeeuw, and D. de Vries, D. 1998. Variable acoustics by wavefield synthesis: A closer look at amplitude effects. In *104th Convention*. Paris, France: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=8468>.
- Spors, S., and J. Ahrens. 2008. A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling. In *125th Convention*. Audio Engineering Society, San Francisco, CA, USA, <http://www.aes.org/e-lib/browse.cfm?elib=14708>.
- Spors, S., and J. Ahrens. 2009. Spatial sampling artifacts of wave field synthesis for the reproduction of virtual point sources. In *126th Convention*. Munich, Germany: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=14940>.
- Spors, S., H. Wierstorf, M. Geier, and J. Ahrens. 2009. Physical and perceptual properties of focused virtual sources in wave field synthesis. In *127th Convention*. New York, NY, USA: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=15109>.
- Spors, S., H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter. 2013. Spatial sound with loudspeakers and its perception: A review of the current state. *Proceedings of the IEEE* 101 (9): 1920–1938. <https://doi.org/10.1109/JPROC.2013.2264784>.
- Start, E. 1997. *Direct sound enhancement by wave field synthesis*. Ph.D. thesis, Delft University of Technology, The Netherlands.
- Suzuki, Y., A. Honda, Y. Iwaya, M. Ohuchi, and S. Sakamoto. 2020. Toward cognitive usage of binaural displays. In *The Technology of Binaural Understanding*, eds. J. Blauert and J. Braasch, 665–695. Chap. 22. Springer and ASA Press.
- Takanen, M., O. Santala, and V. Pulkki. 2014a. Visualization of functional count-comparison-based binaural auditory model output. *Hearing Research* 134: 147–163.
- Takanen, M., H. Wierstorf, V. Pulkki, and A. Raake. 2014b. Evaluation of sound field synthesis techniques with a binaural auditory model. In *55th International Conference: Spatial Audio*. Helsinki, Finland: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=17361>.
- Theile, G. 1980. Üdie Lokalisation im überlagerten Schallfeld (On Localization in the superposed sound field). Ph.D. thesis, Technische Universität Berlin.
- Theile, G. 1986. On the standardization of the frequency response of high-quality studio headphones. *Journal of the Audio Engineering Society* 34 (12): 956–969. <http://www.aes.org/e-lib/browse.cfm?elib=5233>.
- Theile, G. 1991. On the naturalness of two-channel stereo sound. *Journal of the Audio Engineering Society* 39 (10): 761–767. <http://www.aes.org/e-lib/browse.cfm?elib=5963>.
- Thurlow, W.R., and C.E. Jack. 1973. Certain determinants of the ventriloquism effect. *Perceptual and Motor Skills* 36 (3–suppl): 1171–1184. <https://doi.org/10.2466/pms.1973.36.3c.1171>. PMID: 4711968.
- Tucker, A.J., W.L. Martens, G. Dickens, and M.P. Hollier. 2013. Perception of reconstructed sound-fields: The dirty little secret. In *52nd International Conference: Sound Field Control - Engineer-*

- ing and Perception*. Guilford, UK: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=16906>.
- Vilkaitis, A., and B. Wiggins. 2017. WFS and HOA: Simulations and evaluations of planar higher order ambisonic, wave field synthesis and surround hybrid algorithms for lateral spatial reproduction in theatre. In *4th International Conference on Spatial Audio*, Graz, Austria.
- Wang, D., and G.J. Brown. 2006. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New Jersey: Wiley-IEEE Press.
- Wang, L., and S. Kuriki. 2012. Functional cortical mapping of auditory illusion: An fMRI investigation of “scale illusion”. In *2012 5th International Conference on BioMedical Engineering and Informatics*, 117–120, <https://doi.org/10.1109/BMEI.2012.6512934>.
- Warren, R.M. 1983. Auditory illusions and their relation to mechanisms normally enhancing accuracy of perception. *Journal of the Audio Engineering Society* 31 (9): 623–629. <http://www.aes.org/e-lib/browse.cfm?elib=4557>.
- Welch, R.B., and D.H. Warren. 1980. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin* 88 (3): 638–667. <https://doi.org/10.1037/0033-2909.88.3.638>.
- Wierstorf, H. 2014. Perceptual assessment of sound field synthesis. Ph.D. thesis, Technical University of Berlin.
- Wierstorf, H., C. Hohnerlein, S. Spors, and A. Raake. 2014. Coloration in wave field synthesis. In *55th International Conference: Spatial Audio*. Helsinki, Finland: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=17381>.
- Wierstorf, H., A. Raake, M. Geier, and S. Spors 2013. Perception of focused sources in wave field synthesis. *Journal of the Audio Engineering Society* 61 (1/2): 5–16. <http://www.aes.org/e-lib/browse.cfm?elib=16663>.
- Wierstorf, H., A. Raake, and S. Spors. 2012. Localization of a virtual point source within the listening area for wave field synthesis. In *133rd Convention*. San Francisco, CA, USA: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=16485>.
- Wierstorf, H., A. Raake, and S. Spors. 2017. Assessing localization accuracy in sound field synthesis. *Journal of the Acoustical Society of America* 141 (2): 1111–1119. <https://doi.org/10.1121/1.4976061>.
- Wierstorf, H., A. Raake, and S. Spors. 2013. Binaural assessment of multichannel reproduction. In *The Technology of Binaural Listening*, ed. J. Blauert. Chap. 10. Springer and ASA Press.
- Willcocks, M.E.G., and G. Badger. 1983. Surround sound in the eighties: localization and psychoacoustics. In *74th Convention*. New York, NY, USA: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=11734>.
- Wisniewski, M.G., G.D. Romigh, S.M. Kenzig, N. Iyer, B.D. Simpson, E.R. Thompson, and C.D. Rothwell. 2016. Enhanced auditory spatial performance using individualized head-related transfer functions: An event-related potential study. *Journal of the Acoustical Society of America* 140 (6): EL539–EL544. <https://doi.org/10.1121/1.4972301>.
- Wittek, H., S. Kerber, F. Rumsey, and G. Theile. 2004. Spatial perception in wave field synthesis rendered sound fields: distance of real and virtual nearby sources. In *116th Convention*. Berlin, Germany: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=12711>.
- Wittek, H., F. Rumsey, and G. Theile. 2007. Perceptual enhancement of wavefield synthesis by stereophonic means. *Journal of the Audio Engineering Society* 55 (9): 723–751. <http://www.aes.org/e-lib/browse.cfm?elib=14192>.
- Woodward, J.G. 1977. Quadraphony - a review. *Journal of the Audio Engineering Society* 25 (10/11): 843–854. <http://www.aes.org/e-lib/browse.cfm?elib=3315>.
- Zacharov, N., and T.H. Pedersen. 2015. Spatial sound attributes - development of a common lexicon. In *139th Convention*. New York, NY, USA: Audio Engineering Society. <http://www.aes.org/e-lib/browse.cfm?elib=17992>.
- Zatorre, R.J., M. Bouffard, P. Ahad, and P. Belin. 2002. Where is ‘where’ in the human auditory cortex? *Nature Neuroscience* 5 (9): 905–909. <https://doi.org/10.1038/mn904>.

- Zhang, W., P.N. Samarasinghe, H. Chen, and T.D. Abhayapala. 2017. Surround by sound: A review of spatial audio recording and reproduction. *Applied Sciences* 7 (5): 532. <https://doi.org/10.3390/app7050532>.
- Zotter, F., M. Frank, M. Kronlachner, and J.-W. Choi. 2014. Efficient phantom source widening and diffuseness in ambisonics. In *EAA Joint Symposium on Auralization and Ambisonics*, Berlin, Germany.

Creating Auditory Illusions with Binaural Technology



Karlheinz Brandenburg, Florian Klein, Annika Neidhardt, Ulrike Sloma and Stephan Werner

Abstract It is pointed out that beyond reproducing the physically correct sound pressure at the eardrums, more effects play a significant role in the quality of the auditory illusion. In some cases, these can dominate perception and even overcome physical deviations. Perceptual effects like the room-divergence effect, additional visual influences, personalization, pose and position tracking as well as adaptation processes are discussed. These effects are described individually, and the interconnections between them are highlighted. With the results from experiments performed by the authors, the perceptual effects can be quantified. Furthermore, concepts are proposed to optimize reproduction systems with regard to those effects. One example could be a system that adapts to varying listening situations as well as individual listening habits, experience and preference.

1 Introduction

The desire to create a perfect auditory illusion for listeners has been voiced since the invention of technical devices for recording and reproducing sound. Thereby the audio system itself should meet the requirements and expectations of the users regarding immersion (Heeter 1992) and plausibility (Lindau and Weinzierl 2011; Kuhn-Rahloff 2011), or even authenticity (Brinkmann et al. 2017). In other words, the technical system is intended to create an auditory illusion suitable for certain requirements and expectations. In fact, binaural technology has the capability to do so.

As is also addressed in other chapters of this volume (Nicol 2020; Mourjopoulos 2020), creating auditory illusions depends on a variety of cues for the brain. Linear and time-invariant models are not sufficient to model the behavior of the ears and the brain. Auditory illusion can be created via either loudspeaker or headphone reproduction. This chapter concentrates on dynamic binaural synthesis via headphones

K. Brandenburg (✉) · F. Klein · A. Neidhardt · U. Sloma · S. Werner
Electronic Media Technology Group, Technische Universität Ilmenau, Ilmenau, Germany
e-mail: karlheinz.brandenburg@tu-ilmenau.de

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_21

for this purpose. This is the creation of an auditory illusion in a possibly changing environment, such as via tracking the position and orientation of a listener in a room.

It has been shown that with individual binaural recording at the same position in the same room, the majority of the participants could not distinguish between real and simulated sound fields (Brinkmann et al. 2017). Thus, an authentic reproduction can be achieved under special conditions.

Questions: *Is this only possible under exactly these conditions? Which role might context dependent effects have played in this experiment?*

Earlier systems have tried to recreate the exact sound pressure at the ear drums. This is clearly not sufficient, and in some cases not even necessary. Different cues that help or inhibit plausible auditory illusions are

- Correct sound pressure at the ear drum, for example, via measured HRTF or BRIR
- Individual HRTF or BRIR
- Enabling head rotation
- Enable interactive exploration in dynamic orientation and position with tracked self-motion
- Room convergence
- Audiovisual congruence
- Training of listeners to the system (experience).

Questions and related ideas along these lines of thinking have been discussed and validated at several places in this book. The current chapter adds further experimental data on the room-divergence effect, the influence of movements of the listeners in a listening area compared to just sitting, and data on training and adaptation effects.

1.1 Context Dependencies of Binaural Understanding

In previous investigations, many cognitive effects influencing the impression of auditory illusions have been determined and named. Some of these effects are mentioned in the following as examples for context dependencies that need to be considered in the context of plausible binaural understanding.

The well known *Cocktail-Party effect* (Cherry 1953; Bronkhorst 2000) describes attention-based listening. Namely, among a number of concurrent sound sources, persons are able to draw their attention to a specific one, for instance, to a conversation partner or speaker in a crowded room. Perception is always a multimodal process, therefore the visual sense has a high impact on scene understanding. The *Ventriloquism effect* is an example for this (Bertelson and Radeau 1981; Seeber and Fastl 2004). It describes that the perceived location of a sound (the auditory event) is influenced by the visual position of the sound source. A further effect to be mentioned is the *McGurk effect*, (McGurk and MacDonald 1976), which states that the sounds that are perceived can be modified by what is concurrently seen. Further influences towards the perceptive impression are the motivation for listening to a sound source, the individual experiences and expectations regarding how an auditory event sounds.

The following concentrates on selected context dependencies that support the creation of auditory illusion. These dependencies are *congruence of the current listening room and the reproduced room*, the *ability of a listener to adapt to rooms and binaural-synthesis systems*, further, the *interaction of auditory perception and self motion*. These effects are actually hard to measure because they are highly individual, dynamic in time, and the interactions are rather complex. An investigation into them is thus challenging, among other issues, since these effects cannot easily be separated and isolated.

1.2 Outline of This Chapter

Following the introduction in a first section the *room-divergence effect* (RDE) is described from different points of view. Conclusions for room-related binaural synthesis are drawn and evaluated with several experiments using the *direct-to-reverberant energy ratio* (DRR) and *externalization*, a quality feature, as instruments.

The next section then deals with *auditory adaptation* and *training effects*. It is shown, that listeners are able to learn to listen and localize with different head-related transfer functions (HRTFs). Subsequently, the adaptation to room acoustics is assessed. For this reason, the room-divergence effect is exploited.

Auditory-scene exploration by *interactive changes in the listening perspective* is addressed in the subsequent section. These changes may provide additional information with regard to the interpretation of the sound-pressure as captured by the ears. It is discussed which additional acoustical cues are available and potentially have an influence on the interpretation of sounds when listening during walking. This is analyzed for several examples like *distance perception* or *source-directivity estimation*.

In a concluding section it is demonstrated how the previous results can be used and implemented for *synthesis of binaural room-impulse responses*. Already realized and prospective applications are presented. Finally, an outlook points out further analysis tasks and questions that are relevant for the creation of auditory illusions by mean of binaural synthesis.

2 Room-Divergence Effect

The room-divergence effect (RDE) describes what happens in terms of auditory illusion when the room acoustics of the recording room and the room acoustics of the synthesized room differ in terms of spatial auditory perception. If such divergence exists, the perceived externalization of the auditory event is reduced. The reason for this effect lies in a cognitive dissonance between an expected and a currently perceived auditory event.

A fundamental approach to explain this effect is based on the auditory *precedence effect* (Wallach et al. 1949). This effect describes that the first wavefront arriving from a sound source at a listener's ears is prioritized when forming the position of the resulting auditory event. Sound waves arriving after the first wavefront are perceptually assigned to this position as long as the time difference of the first wavefront and succeeding ones is shorter than a specific *echo threshold*. For delays larger than this threshold, distinct echoes are perceived after the initial auditory event—possibly in different directions. Significant extensions of the precedence effect are those from Clifton (1987) and Litovsky et al. (1999). In their experiments, the pattern of direct sound and reflected sound were spatially modified. A change of the pattern leads to a reduction of the echo threshold—commonly referred to as the *Clifton effect*—and the precedence effect breaks down. The change sets a new precedence, yet, when the old situation is recovered, the old precedence effect is still apparent for some seconds—compare (Blauert 1997) and, in particular, the effect of room learning, as has been investigated in detail by Seeber and Clapp (2020), this volume.

The schemata—for the term *schema* compare Sutojo et al. (2020), this volume—as stored in the auditory system for recognizing space and audio scenes may not match those derived from the synthesis. If in such a case the deviations are sufficiently large, the cognitive system is no longer able to reach a perceptive fusion between the synthesized room and the listening room. The assimilation of what is currently perceived onto a stored schema/pattern then fails. As is explained in the following, this effect is hypothesized to be triggered by room-acoustic divergences between synthesized rooms and the listening room. An auditory-visual divergence seems to intensify this effect but is, in itself, not sufficient for a conclusive explanation.

2.1 *Effect on Externalization*

The term *Externalization* describes the perception of the location of an auditory event outside the head. The counterpart to this is the *In-Head-Localization* (IHL). By definition, IHL takes place when the auditory event is positioned within the head. The boundary surface of the head is thus clearly defined as the boundary between IHL and externalization. The quality characteristic externalization is assigned a bipolar characteristic value. The perception of auditory events outside the head is seen as an essential quality feature of a binaural headphone system to create a plausible spatial auditory illusion.

From investigations by Toole (1970) and Plenge (1972) it is known that the effect of the IHL is not necessarily dependent on the use of a headphone system. Toole was able to show in his experiments that IHL also takes place when using loudspeakers in anechoic environments (Toole 1970). The test subjects listen to audio signals from single loudspeakers in front of and behind the person as well as from up to four loudspeakers around the test subject. There is simultaneous sound from one, two or four loudspeakers with noise signals of different bandwidth. It shows that when the audio signals are presented via two and four loudspeakers, a relative probability of

externalization of the auditory events of less than 20% is achieved. When presenting audio signals from a speaker behind the subject, externalization values of $\approx 80\%$ are achieved for broadband white noise. The presentation of audio signals from a frontal loudspeaker leads to values of $\approx 65\%$ for broadband white noise.

Toole also investigated the influence of minor head movements to identify the type of auditory event. For this purpose, he minimized the possible head movements of the test listeners (Toole 1970). Excluding relative changes in the direction of sound incidence reveals similar assessments of externalization as in the previous experiment. Toole concludes that large searching head deflections lead to a correct localization of the sound events. On the other hand, small deflections do not seem to be sufficient for an increase in the externalization of auditory events (Toole 1970).

In a further study on the emergence of IHL by Plenge (1972) the hypothesis is raised of that the IHL arises from a lack of assimilation or an inadequate learning process. The learning process includes instant learning of characteristics of the sound source and the listening room. Experiments were conducted (a), “[...] to create a smooth transition between out-of-head localization and IHL [...]” (Plenge 1972), (b), to perform a comparison of head-related electroacoustic storage with the original signal (Plenge 1972). In the experiments, the above mentioned disturbances in localization were artificially caused by preventing the learning of sound source and room characteristics (Plenge 1972).

Experiment (a) showed that a smooth transition between externalization and an IHL cannot clearly be established in a low-reflection room (Plenge 1972). Even small changes in the test signals lead either to IHL or to the perception of externalization. Based on the results of this experiment, the externalization can be seen as a bipolar quality feature. Experiment (b) dealt with the comparison of loudspeaker reproduction with the binaural synthesis of the same loudspeaker configuration, but simulated via headphones. Results show that occurrence of IHL is “*independent of headphone reproduction*” (Plenge 1972). As a conclusion, Plenge states that the IHL arises when the ear signals “*cannot and must not be assigned to a real source outside the head—they must not be confusable with any real source—must be accommodated in the only remaining place where no sound source can occur, namely in the head*” (Plenge 1972).

In the first part of experiment (c), a synthetic sound field of a concert hall was generated in a low-reflection room. Direct sound, early reflections, and diffuse reverberation are approximated by loudspeaker reproduction. As a test signal, the reverberation-free reproduction of a talker via a loudspeaker in the direction of view in the room is used. The test listeners were supposed to assess whether the auditory event is inside the head, at the surface of the head, or out of the head. In only one out of 68 evaluations, the talker was assessed as being external (Plenge 1972). In the second part of the experiment, a sound field of a concert hall or an outdoor recording (such as street noise) recorded over a dummy head is binaurally presented to the test persons via headphones. The test signal is then reproduced unchanged via a loudspeaker, but now contains either reverberation-free or reverberant speech or music signals. Plenge states that an inside-the-head or at-the-head localization occurs when there is a “*missing, inadequate or incorrect sound source and field knowledge and/or*

the signals and, thus, the stimuli are such that they cannot be assigned to any stimulus pattern contained in the long-term memory” (Plenge 1972). These results suggest that the quality feature externalization is influenced by the context of the playback and the listening situations. This is indeed the *room-divergence effect* (RDE).

2.2 Room Divergence as a Visual Effect

A possible approach to understand the RDE is that the reduction of externalization of auditory events is due, among other things, to a mismatch of the visual impression of the listening room and the binaurally synthesized room. Further influences as having been identified are individual listener characteristics and the direction of sound incidence.

Udesen et al. (2015) present an experiment where a room, similar to a living room with loudspeakers in it, is binaurally auralized. The test conditions include a visual variation of the listening room. The (virtual) audio signal of the living room is kept constant. This results in the conditions for real living rooms (RLR) with audio presentation via real speakers in the room and virtual living rooms (VLR) with binaural-synthesized speakers. A hall is used as a second room. In this hall, the positions of the speakers to be synthesized are visible as dummies (VHWS, virtual hall with speakers) or not visible (VHWOS, virtual hall without speakers). The audio signal is identical to the condition VLR. According to the definition of the room divergence effect, the conditions RLR and VLR represent room-convergent situations. The conditions VHWS and VHWOS are room-divergent situations.

The assessments of the externalization of auditory events show that the room-convergent conditions RLR and VLR achieve high externalization values. The room-divergent conditions VHWS and VHWOS each achieve significantly lower values than the condition VLR. It is concluded that “The only parameter that was changed between the virtual test environments was the visual stimuli during the tests. In the VLR test environment, the reverberation corresponded to the visual impression of the room, while in the VHWS environment there was a discrepancy between the virtual reverberance and the visual impression of the room” (Udesen et al. 2015). No significant difference is determined between conditions VHWS and VHWOS. Thus the conclusion is “[...] *that the visual impression of space influences the externalization of sound more than visual presence of the speakers*” (Udesen et al. 2015).

In the view of the notion of the current authors with regard to the room-divergence effect, this statement must be put into perspective. Rather than the claim that the difference in externalization is due to the audio synthesis, namely, as a result of mismatch of the learned in the cognitive system and the currently experienced room acoustics. The presence of appropriate visual stimuli increases the degree of externalization but does not explain the room-divergence effect. To confirm this statement the current authors have performed experiments of their own (Werner and Klein 2014; Werner et al. 2016). The studies of Udesen et al. (2015) also found an influence of the individual listener and the direction of sound incidence. Sound sources from the azimuth

0° externalize significantly worse than sources from 180° and 90°. This effect is also addressed and confirmed in the current authors' own investigations (Werner and Klein 2014; Werner et al. 2016).

2.3 Room-Divergence as a Combined Auditory and Visual Effect

A further study on the room divergence effect has been published by Gil-Carvajal et al. (2016). In this investigation loudspeaker positions from different rooms and directions are binaurally synthesized. The test conditions included the separate and common visual and auditory presentation of different rooms. A reference room ($T60 = 0.4\text{ s}$, $V = 99\text{ m}^3$), a small and reverberant room ($T60 = 2.8\text{ s}$, $V = 43\text{ m}^3$) and a large and nearly anechoic room ($T60 < 0.01\text{ s}$, $V = 330\text{ m}^3$) was used. Sound sources to be synthesized were located at 1.5 m distance in the azimuth 0°, 60°, 90°, 180°, 210°, and 270°. Individual BRIRs from the reference room were used for binaural synthesis. The listening room was always the reference room. This allowed for visual and/or acoustic room-divergent and room-convergent test conditions. The evaluation of the externalization is rated on a 6-point distance scale from 0 ...in-head, to 4 ...listening event at the speaker distance and, 5 ...listening-event distance larger than speaker distance.

The results show that an effect according to room convergence or divergence occurs for the evaluations of conditions with visual and auditory room characteristics. In the case of room-divergent conditions, the auditory events are rated as been closer to the room-convergent reference condition (Gil-Carvajal et al. 2016). The largest effects are found between the reference condition and the small reverberant room. There are also significant differences between room-convergent and room-divergent conditions in distance evaluations of test conditions with purely auditory characteristics. However, the observation of purely visual characteristics and the resulting room convergence or divergence does not result in any significant differences (Gil-Carvajal et al. 2016).

Gil-Carvajal et al. (2016) conclude that auditory characteristics have a greater influence on the externalization of auditory events than visual characteristics. They note that visual characteristics should be differentiated between room-related characteristics and sound-source-related characteristics. Gil-Carvajal et al. also find that the room divergence effect is stronger for front and rear sound-source positions than for side positions.

2.4 Room-Divergence from the Current Authors' Point of View

In our own investigations, in particular, sound source related visual characteristics (visibility of the loudspeakers and the room) and the influence of the individualization of binaural synthesis, is examined in relation to the room-divergence effect. Furthermore, a direct measurement of the externalization of auditory events without using a distance evaluation was used. The room-divergence effect was investigated for two rooms as synthesized room and listening room. This resulted in a complete test design with all four possibilities of room divergence and room convergence. A variety of experiments investigated the RDE in different ways and with different emphases (Werner and Siegel 2011; Werner et al. 2013, 2016; Werner and Klein 2014). The results show a high agreement regarding the RDE.

The object of investigation for the experiments is an acoustic divergence between the resynthesized room and the listening room. A listening laboratory (abbr. LL; Rec. ITU-R BS.1116-1, $V = 179 \text{ m}^3$, $RT60 = 0.34 \text{ s}$) and an empty seminar room (abbr. SR; $V = 182 \text{ m}^3$, $RT60 = 2.0 \text{ s}$) were used for the listening test and the measurements of the BRIRs. The experiment was conducted at the same recording positions in each room to evaluate the influence of the listening situation.

Discrete sound source directions were considered in the tests. Loudspeakers were used to measure the BRIRs for each position. The investigated azimuths are 0° , 90° , 120° , 180° , and 330° (clockwise orientation). The distance from the loudspeakers to the listening point was $\approx 2.2 \text{ m}$. The height of the source position was $\approx 1.3 \text{ m}$ (ear position of a sitting person). The BRIRs for each position and for each test person were recorded individually in the two rooms. The recording position was the same as the listening position in the tests.

The individual BRIRs and Headphone Transfer Functions (HPTFs) of the test persons for both rooms and sound-source directions were recorded. Furthermore, the BRIRs and HPTFs of a KEMAR head-and-torso simulator (45BA) were recorded. Both the individual and artificial BRIRs were then used to create the binaural test stimuli. The panel of test persons was randomly divided into two groups with or without the presence of visual cues during the experiment. For the first group, the illumination of the listening rooms was minimized to nearly complete darkness, such that the test persons should not have any visual impression or additional visual cues regarding the listening rooms. In contrast, the test persons in the second group were placed in the illuminated listening rooms to provide additional visual cues.

The experiment thus assessed the perceived externalization at different synthesis conditions. An individualized and artificial binaural synthesis of a single loudspeaker with different source directions was used. The test conditions were the four combinations of synthesized room and listening room, that is, LL in LL, LL in SR, SR in LL, SR in SR. After the individual and artificial BRIRs, a set of free-field KEMAR head-related impulse responses was used to synthesize test stimuli. Free-field stimuli were used as anchors with a low spatial quality. Prior to the test, the test persons were familiarized with the listening conditions and the user interface by listening

to individualized and artificial synthesized sounds in divergent and congruent room conditions, to the anchor signals, and to the playback of the real loudspeakers in the listening room. It was intended that the internal reference of the test persons should not shift explicitly to a congruent or divergent room condition. Twenty-three test persons participated in the evaluation. Dummy loudspeakers were placed around the listeners in steps of 30° to provide additional visual cues for the group in the illuminated conditions. The test persons rated externalization on a scale with the categories 1, ...in-head, 2, ...external, but close to the head, and 3, ...external, in the room. The data analysis was based on an externalization index, namely, the ratio between the number of ratings for external, in the room and the overall number of ratings. A more detailed discussion regarding the used scale can be found in (Werner et al. 2016).

The context parameters *room divergence*, *visual influences*, and *individualization* were analyzed. To analyze the statistical independence of the context parameters, a Chi^2 -test was performed on each of the context parameters. No significant correlation between the parameters was found ($p < 0.05$; Chi^2 -test on independence). This remains valid for room divergence and additional visual cues for the synthesized room as well as for the listening room, the personalization, and additional visual cues.

Figure 1 shows an excerpt of the rating results. The perceived externalization is dependent on the presented source direction. Less externalization is measured for virtual audio objects placed at directions with expected localization inaccuracies. These directions are 0° with front-back-confusions and 180° with back-front-confusion. The lowest externalization is observed for the anchors (free-field). Significant increases of externalization (Fisher's exact test) is observed for room congruence compared with divergent listening conditions. This effect is, for example, visible for the 0° -direction and for the individualized synthesis of the seminar room, that is, SR(I) in SR compared to SR(I) in LL ($p < 0.01$ with an effect size as indicated by an odds ratio of 4.0). A similar increase of externalization is also visible for individualized synthesis, that is, LL(I) in LL compared to LL(I) in SR ($p < 0.01$ and effect size as indicated by an odds ratio of 3.0). An odds ratio higher than one indicates a positive impact of the convergent room condition in relation to the divergent condition. A similar tendency but at a lower probability value (p-value) is visible for synthesis using artificial BRIRs and the other directions.

The analysis of externalization between the groups with and without visual cues shows significant increases, especially for the 0° and 180° directions (Fisher's exact test). No clear tendencies with respect to the personalization method and combination of the listening room and synthesized room are detectable. An increase of externalization of 16% is measured as the mean value across all source directions, personalization methods and room conditions. A higher increase of perceived externalization of 26% is observed for the 180° direction, independent of the other test conditions—with a probability value of difference of at least $p < 0.1$ in six of eight test conditions. This is an unexpected effect because the test persons were instructed to keep still during the experiment and were not able to see the rear positions.

The externalization increases when individualized binaural synthesis is used as compared to using artificial BRIRs. Highly significant differences are visible for the

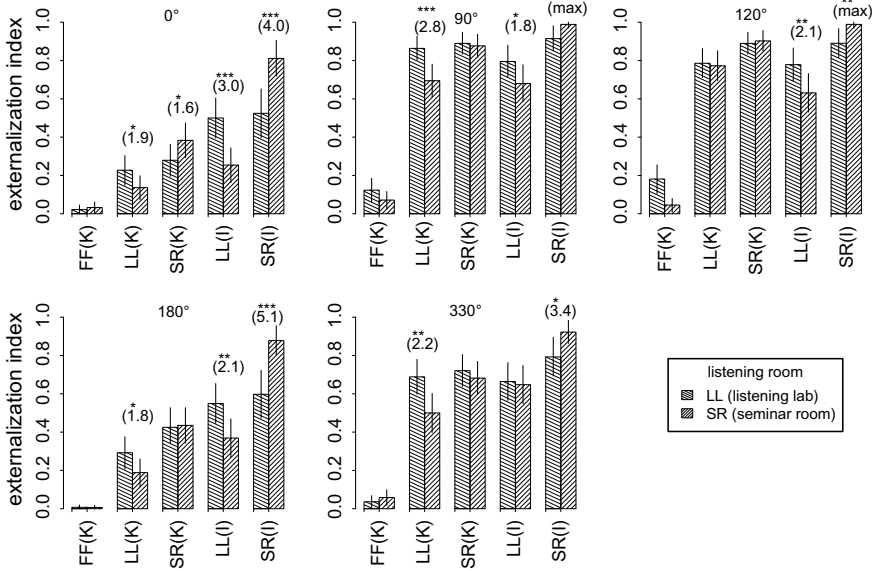


Fig. 1 Externalization indices with a 95% confidence interval for combinations of listening room and synthesized room, individualization of binaural synthesis, and for summarized ratings with and without visual cues. SR ...seminar room, LL ...listening lab, FF ...free field, K ...artificial-head BRIRs, I ...individual BRIRs, *** probability value of difference at $p < 0.01$, ** probability value of difference at $p < 0.05$, * probability value of difference at $p < 0.1$. Numbers in brackets note the effect sizes in terms of odds ratio. Figure after Werner et al. (2016)

0° and 180° directions ($p < 0.01$) with effect sizes in odds ratio from 34 up to 10; (Fishers exact test). The effect size of the increase of externalization seems to be higher for the personalization method compared to the room-divergence effect. As an example, an effect size of ≈ 5 is observed for the synthesis of the seminar room using individual BRIRs and for the 180° direction. An effect size of 10 is reached for the synthesis SR in SR for the same direction, dependent on personalization. Similar tendencies are also observable for the 0° direction and other room combinations. Furthermore, higher effect sizes are seen for congruent room conditions compared to divergent room conditions, especially for the 0° and 180° directions. This gives a hint on a negative correlation between perceived externalization and the occurrence of localization errors, such as front-back or back-front confusions. Room divergence can cause localization errors and therefore less externalization. A more detailed analysis of this effect can be found in Werner et al. (2016). Yet, further investigations into this correlation are obviously needed.

Overall, there is some evidence that the observed effect of the different room combinations is an audition-based context-dependent quality parameter. This parameter has an influence on the quality feature externalization. The influence is a result of adaptation to and expectation of the room acoustics of the listening room.

3 Auditory Adaptation Effects

Auditory adaptation effects are well-known in a broad range of research areas. In neuroscience, sound-localization experiments are studied to obtain information on the neural processes related to adaptation. A recent overview is given by Keating and King (2015). In the field of hearing-aid treatment auditory adaptation effects are utilized for training of speech intelligibility and for the identification of everyday sounds. In hearing research, processes of auditory adaptation are observed for frequency discrimination tasks. In listening tests on audio quality, listeners typically get trained in order to rate specific quality attributes more reliably or to detect specific coding errors.

It is well-known that the ability for spatial hearing is not only based on signal-driven processing but also on listener experience and expectations. Expectations about sound sources or room characteristics serve as an internal reference for the listener in order to rate the perceived quality—compare Raake and Wierstorf (2020), this volume. These expectations can change depending on prior sound exposure. When these mechanisms apply to spatial hearing, it means that listeners are probably able to learn how to interpret spatial cues such as head-related transfer functions or room reflections.

Such adaptation effects are rarely taken into account during the development of binaural-synthesis systems or other spatial-audio-reproduction techniques. A more detailed understanding is necessary to determine the relevance of adaptation processes regarding the plausibility of a virtual acoustic scene. Furthermore, research on auditory-adaptation effects enables new technology advances by identifying weaknesses of technical components that could be compensated or exaggerated by these effects. The following sections briefly describe two cases, where auditory adaptation effects have an influence on the quality rating.

3.1 *Adaptation to Artificial Localization Cues*

To investigate the process of auditory adaptation, ear signals are artificially distorted to introduce perceptual errors like increased localization errors or degradation of externalization. In earlier studies, ear molds were inserted monaurally or binaurally as described, for example, by Hofman et al. (1998). In more recent publications, researchers have distorted the ear signals by altering the head-related transfer function (HRTF) in binaural-synthesis systems. Depending on the method, alteration of the spatial cues were realized, for instance, by frequency warping (Majdak and Labak 2013), using HRTFs with a different degree of personalization (Parseihian and Katz 2012) or by applying HRTFs measured with an artificial head (Zahorik et al. 2006; Mendonça et al. 2013; Mendonça 2014). In the next step, participants are trained to a new set of spatial cues employed by the altered HRTFs. When it comes to training methods, a wide variety of possibilities can be found in literature.

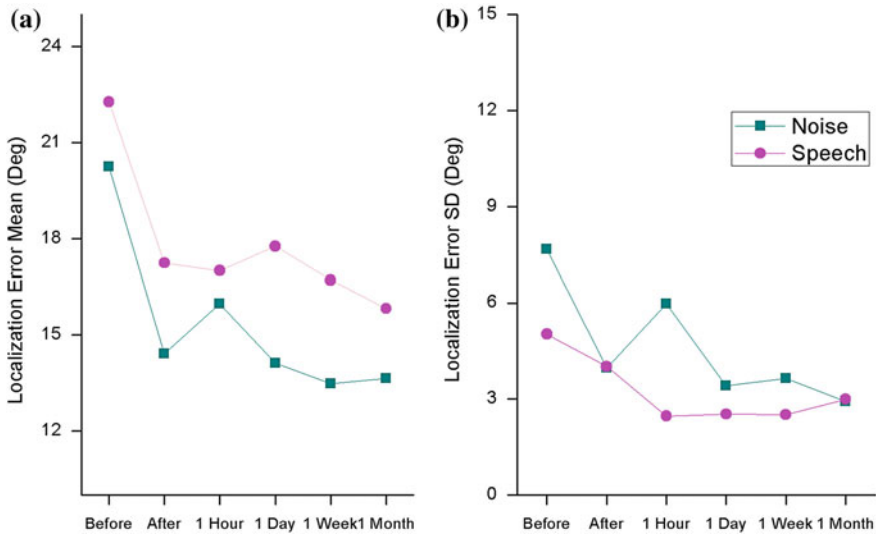


Fig. 2 Results for the post training tests of the azimuth experiment. **a** Shows the average azimuth localization error, and **b** the respective standard deviations (SD) for both stimuli (Mendonça et al. 2013)—reproduced courtesy of C. Mendonça

Basically, there are two common types of training, that is, sound exposure and training with feedback. The type of feedback defines which reference information is provided for training. Acoustical, visual, verbal or tactile information about the real sound direction can serve as feedback. A proprioceptive detailed overview of publications related to auditory training recently has been published by Mendonça (2014). Different types of training, auditory stimuli, types of cue changes, and durations of training are considered for comparison.

The training procedure of Mendonça et al. (2013) consisted of an azimuth and an elevation experiment. In both experiments the listeners actively learned to identify the position of several sound sources at first. Consequently, they conducted a training session with visual positional feedback. Listeners were trained to listen with anechoic non-individual HRTFs with an overall training time of about 10–20 min. Afterwards, post-tests were conducted at different time intervals. The results showed a decrease of localization error for azimuth and elevation angle. Figure 2 shows the training effect for the azimuth-localization error. The error was reduced immediately after training and remained low in subsequent tests.

Parseihian and Katz (2012) trained listeners in a virtual game-like scenario where they could navigate using a trackball. During the training, positional feedback was given acoustically and by proprioceptive information. The training took 36 min and was split into three sessions. Participants were trained group-wise according to the degree of individualization used—individual HRTF, good fitting HRTF, and bad fitting HRTF. The training mainly decreased the polar-angle error which is related to the spectral cues. Additionally, they found hints that the adaptation time depends

on the degree of cue change. In an experiment by Klein and Werner (2016), visual feedback was provided for the training on artificial localization cues—binaural room impulse response recorded with a KEMAR artificial head. Similar to the results of Parseihian and Katz (2012), an improvement was mainly found for polar angles. The training environment in Majdak and Labak (2013) consisted of a head-mounted display that places the listener in a spatial virtual visual environment and provides visual positional feedback. HRTFs for the auditory stimuli were either band-limited or spectrally warped. In comparison to other studies, the training time was long with training sessions of 2 hours per day for 21 days. Their results showed a decrease of quadrant errors, that is, including front-back or up-down confusions, depending on the amount of training. Similar to Parseihian and Katz (2012) they found an interconnection between the amount of cue change and the duration of the adaptation process. In general, experiments with visual feedback during training show a faster adaptation process than experiments with other feedback information.

3.2 *Adaptations to Room Acoustics*

As outlined in Seeber and Clapp (2020), this volume, mechanisms of adaptation (or abstraction) also apply for reflections in a room or even perceptual models of a room geometry. The re-calibration to different reflection patterns or rooms has shown to affect localization and speech understanding. The “break-down” of a perceptual room model may also be the reason for the described room-divergence effect, namely, the perceptual model of the actual listening room conflict with the room presented over headphones. This obviously influences the perceived externalization negatively. Further research should investigate the possibility of increasing the externalization by controlling the listeners’ expectations in suitable training sessions. Similar to the build-up mechanisms of the precedence effect, this could also be the case for more complex room scenarios.

To provoke the room-divergence effect in a current listening test (Klein et al. 2017b), two rooms of similar size but strongly differing reverberation time and direct-to-reverberant ratios were chosen. The basic concept of the listening test was to randomly separate the participants into two groups, each trained to different room acoustics with individually tailored stimuli. After the training, both groups were faced with a familiar and unfamiliar room condition to measure how the training sessions would influence the externalization ratings. Individual BRIRs were measured in both rooms for 31 participants. The training was designed as a simple localization task accompanied with a judgment on the perceived level of externalization. Next to the rating task, visual feedback was provided at the correct source position by visually highlighting a loudspeaker model. The test task was different to the training task. The participants were asked to rate their personal level of perceived externalization in a single-stimulus test design on a three-level scale of 1, ...in-head, 2, ...near-the-head, and, 3, ...outside-the head.

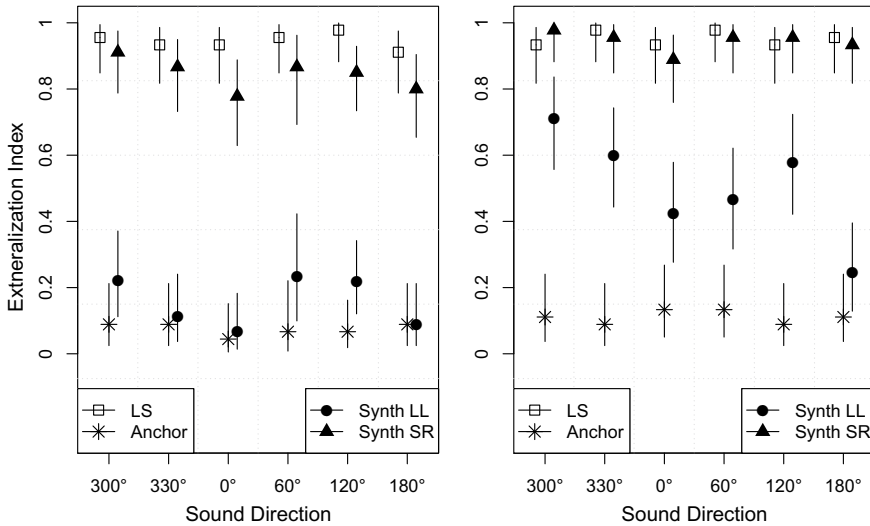


Fig. 3 Results for the externalization ratings of **left**, the convergent group and, **right**, the divergent group. Ratings are separated according to the direction of the presented sound. **Synth SR**, synthesis of the actual listening room, **LS**, real loudspeakers in the listening room, **Synth LL**, synthesis of acoustically dry listening room (Klein et al. 2017b)

Figure 3 shows an excerpt of the test results. The listening tests were conducted in the SR room. The “convergent group” was trained to the SR room (reverberation time at 1 kHz: 2 s) and the “divergent group” to a less reverberant LL room (reverberation time at 1 kHz: 0.339 s). The stimuli “LS” corresponded to actual loudspeakers in the SR room. The low-quality anchor was measured with an omnidirectional microphone in both rooms and aimed to provoke in-head localization. Both groups rated the actual room (LS and Synth SR) high regarding to the externalization, but the rating of room LL is different for each group. Former listening tests have shown that the room-divergence effect is particularly strong when the synthesis of an acoustically dry room is presented in a reverberant room. This was confirmed by the difference between the signals Synth LL and Synth SR of both groups. Additionally, the result of this listening test showed, that the perceived externalization can shift according to the previous listening experience, that is, the particular training session. In other words, the room-divergence effect highly depends on the listeners’ experience in the listening rooms.

For practical application, this means on the one hand that physically correct ear signals cannot guarantee perfect externalization because the basic principle of creating virtual rooms in real rooms may violate the expectations about the room. On the other hand, adaptation to new acoustic environments is obviously possible and Keen and Freyman (2009) state that this process is probably very fast. Therefore, it is likely that the impact of the room-divergence effect depends on the actual application. Video games or movies that immerse the listener in a virtual acoustic environment

provide the possibility to adapt to this environment. However, augmented acoustic applications always present virtual acoustics on top of the real acoustic environment. Any acoustical mismatch is likely to produce confusion and will lead to in-head-localization. Future research has to further investigate the role of adaptation in such acoustic scenarios.

4 Listener Movement and Exploration

Humans use head movements to localize sound sources with a higher accuracy (Blauert 1997; Thurlow and Runge 1967; Thurlow et al. 1967). In the past it was shown, that providing the option of interactive head rotation to the listener in a binaural reproduction improves externalization (Brimijoin et al. 2013; Stitt et al. 2016; Hendrickx et al. 2017), reduces front-back-confusion (Begault and Wenzel 2001) and supports the localization accuracy (McAnally and Martin 2014; Mackensen 2004). Furthermore, when listeners evaluated the timbre of the sound of a source, the range of head motion was relatively small compared to the movements that happen when listener envelopment and source width (Kim et al. 2007) are evaluated. This indicates that head movements are beneficial when judging the spatial impression of virtual auditory scenes.

Schymura et al. (2016) proposed an extension of a binaural listening model, that considers head rotation. The model was investigated by use of a machine-hearing system that makes use of different head motion strategies. The consideration of dynamic acoustic information led to results much closer to real-world observations. This indicates the relevance of dynamic acoustical cues in binaural understanding. Also, it suggests an interaction of self-motion and auditory understanding.

In this section, an overview is provided of research conducted with regard to interactive listener translation. As a consequence of our literature review in combination with observations from own experiments of the current authors, the following hypothesis has been set up.

Hypothesis: *When listening to a scene without movements, after an exploration phase the brain interprets the heard sound in a different way than before the exploration*

The following questions may be raised at this point.

- Is there a benefit from additional information provided by different listening perspectives?
- Is there a useful contribution from an interaction between human listening and self-motion?
- Are there further cues that could be helpful?
- Are there also disadvantages or distortions of the auditory perception that occur in the case of active self-translation?

In the following, aspects of active listener-translation are discussed that may have the potency of affecting the interpretation of the sound-pressure signals—and the

question of how this interpretation might finally contribute to an understanding of the auditory scene.

Carlile and Leung (2016) reviewed a selection of studies on the perception of auditory motion with regard to translation. This includes the motion of sound sources observed by a static listener as well as movements of the listener, for instance, by walking or moving in a wheelchair. The authors point out that, probably, in everyday life, the majority of motions that are experienced are caused by self-motion. Nevertheless, prior studies focused mainly on the perception of auditory motion while listening from a fixed position. Actually, in contrast to auditory perception, much more research has been conducted with regard to the interaction of self-motion and visual cues on motion. The following subsection provides some insight into this topic.

4.1 Interaction of Vision and Self-motion

People who are sitting on a train while watching another train moving often have the impression that their own train is moving. Obviously, visual information can induce a perception of motion (*vection*). The interaction between visual sensory input and the sense of self-motion has been subject to research for more than a century. Durgin (2009) provides a review. Here, only a short summary is given.

The results of a series of experiments suggest that there is a relationship between the visual flow during walking, the vestibular stimulation, and the actions involved in walking, which enables the listener to predict one of them from the other two running 20 s long on a treadmill, a case of self-motion without the expected visual flow, temporarily affects the perception of the distance traveled during blind walking. As a result, people walk too far, when approaching an object seen before moving (Durgin et al. 2005). In contrast, without running on a treadmill, people are quite accurate in this task.

Furthermore, Durgin (2009) examined the theory, that “*in the control of action, perceptual precision (the fineness of discrimination among actual values of a variable) is more important than the perceptual accuracy (direct correspondence between the perceived and actual value of a variable)*”. An example is that looking through a prism that causes a localization offset (Harris 1980). Yet, after adaptation to the prism glasses, a person can hammer a nail despite the given offset. It is only necessary to align the position of the hand holding the hammer with that of the nail. In another study, it was observed that the perception of visual flow is distorted during walking. It appears to be slower than without walking.

In contrast to vision, the interaction between auditory perception and self-motion has not received much attention yet. Vection induced by auditory input only is usually much weaker than vection caused by visual input (Väljamäe 2009; Väljamäe et al. 2005). This suggests that the interaction is not as strong in audition as it is in vision. Nevertheless, there is still some impact of self-motion on auditory perception. For vision, Durgin (2009) developed the hypothesis, that “*rather than emphasizing*

the need for accurate absolute metrics for action [...] the precision of the relative metrics of perception and motor action are much more important". Could this apply to auditory perception as well? Actually, it might provide an explanation for spontaneous exploration movements like walking around. The following section discusses this issue in more detail.

4.2 Relevance of Position Changes for Auditory Perception

Martens and Kim (2009) conducted a study with a special binaural listening instrument that interchanges the ear signals between left and right. This also reversed the listeners' impression of back and front as well as up and down for a moving listener. In an additional experiment, Martens et al. (2009) compared the interchanged mode with the non-interchanged mode in a *standing still* and a *walking* condition. In the standing still condition, both listening modes resulted in similar results in an up-down-discrimination task. In contrast, in the walking condition, the participant had a reversed up-and-down perception in the interchanged mode, but not in the normal-hearing mode. This suggests that the dynamic cues resulting from walking dominate the spectral directional cues when estimating the source height.

In the case of walking, the reversal of the localization of left and right as well as front and back led to a "*Phantom Walker illusion*" (Martens et al. 2011). In other words, the sound source was not perceived as stationary during an approaching motion. Instead, "*the sound was invariably heard to be approaching [them] from behind and the voice of the illusory 'Phantom Walker' overtook listeners as they passed by the physically stationary source*".

In contrast, Macpherson (2011) observed that spectral cues dominated the dynamic cues in cases of head rotation only in virtual auditory space. In this study, flat-spectrum noise of various bandwidth was used, while Martens et al. (2011) used speech, where the energy is usually higher at low frequencies. This might explain the different observations.

Martinson and Schultz (2006) studied the localization of a static sound source in a noisy environment with a moving robot equipped with a microphone array. An approach based on evidence grids made use of the additional information provided by the changes of the recording perspective. This shows that the consideration of dynamic acoustical cues due to position changes can be beneficial in algorithmic analysis. Does the brain make use of those cues as well? The following paragraphs discuss selected aspects of auditory scene perception, which may be understood better when considering that listeners make use of active position changes while exploring the scene.

Distance Perception

No effect of head rotation on distance perception was found in the real environment (Simpson and Stanton 1973). Kearney et al. (2015) studied the same question in a virtual acoustic environment but found no significant impact either. In vision,

however, motion-parallax effects and the time-to-contact are known to support the perception of depth and the estimation of the distance to objects in the scene (Wexler and van Boxtel 2005; Rogers and Graham 1979). To make use of those dynamic cues, certain conditions need to be fulfilled, for example, an efficient velocity (Hayashibe 1996). Shaw et al. (1991) and Guski (1992) proposed the theory that the momentary change of intensity can serve as a dynamic cue and provide additional information. This effect is known as *Acoustic τ* , in analogy to the optical τ . The variable τ denotes the time-to-contact or the time-to-collision when moving towards an object.

Ashmead et al. (1995) observed, that an approaching motion towards a loudspeaker affected the estimated egocentric distance. The participants were blindfolded and had to report the distance by walking to the perceived location. Distances between 5 m and 19 m were tested. Significant differences could be found between listening from a static position, listening from two different static positions and listening during walking. The authors explain their observation with the acoustic τ . Speigle and Loomis (1993) conducted a similar experiment but with distances of 2, 4, and 6 m. Furthermore, sources at different azimuths (0° , 30° , 60° , 90°) with respect to the “origin” (listening position in the static listening condition) and to the direction of movement were added. Thus, in those cases (30° , 60° , 90°), additionally to the acoustic τ , motion parallax effects may provide dynamic acoustical cues. In this experiment, the sound level of the source was not varied, in contrast to Ashmead et al. (1995). The case of listening from a fixed position was compared with two dynamic conditions in which the participants had to walk 2 or 4 m towards the 0° source position to arrive at the position of static listening.

In both dynamic conditions, a tendency could be observed that the overestimation of close distances is smaller, suggesting a higher accuracy of the perceived distances. However, it remains open, whether this applies to the expected underestimation of farther distances as well. Yet, these cases had not been tested in the experiment. Furthermore, the perception of direction was obviously affected in the two dynamic conditions. The authors explained the effect with a practical issue in the blind walking condition, namely, overshooting the “origin” after walking by 62 and 76 cm in average. No significant differences between the distance estimation for the 0° direction and the other directions were found. Thus, this experiment does not show any impact of a motion parallax effect.

Rosenblum et al. (2000) studied the distance perception of a wall in an echolocation scenario. In the moving condition, the estimations were slightly more accurate than in the standing still condition. The authors assume, that the acoustic τ could be the explanation.

Genzel et al. (2018) claim to provide psychophysical evidence for the auditory motion-parallax effect. Listeners had to distinguish whether a high-pitched sound source was closer or farther away than a low-pitched source. When the difference in distance was only 16 cm the participants could not solve the task without motion. Also for higher differences, the results with motion were significantly better. Thus it may be concluded that the participants made use of “time-variant binaural perceptual cues associated with motion” in a distance segregation task. The authors put lots of effort into the minimization and elimination of the known acoustical cues for distance estimation, such as sound level or spectral cues. However, the potential motion-

parallax effects were not separated from disparity. Thus, it is not clear, whether the observed difference is caused by dynamic effects or only by listening from additional perspectives.

However, besides those few experiments, there is a lack of further studies confirming the effects of the acoustic τ or of auditory-motion parallax. Zahorik et al. (2005) summarize that dynamic cues play a minor role in auditory distance estimation. For the case of listener translation, one explanation might be the relatively low speed of walking. During fast position changes, potential effects of dynamic cues may be stronger and cause clearly measurable differences—compare, for instance, Störig and Pörschmann (2013).

Estimation of Source Directivity

The directivity of a sound source describes the frequency-dependent propagation of radiated sound beams as a function of the direction of radiation. This parameter is important to understand the characteristics of the sound source and of resulting reflections in a room. In a real environment, a person is able to walk through the room and around the sound source. The characteristics of the heard sound vary according to the source directivity. For the creation of a plausible virtual environment over headphones with changes of the listening position being allowed, these differences need to be considered. However, the required level of detail is not understood. Is source directivity, potentially only approximated, noticed at all?

The influences of different sound-source directivities on the perception of real and virtual environments have been addressed in several studies. Martin et al. (2007) showed that a variation of the source directivity leads to differences in the measurements of room acoustic parameters. There is a significant influence, especially in the high frequencies. Wang and Vigeant (2008) compared objective and subjective measures of an omnidirectional, a realistic, and an extremely focussed sound source. They used just noticeable differences (JNDs) of the room acoustics parameters clarity and reverberation time to predict whether listeners are able to perceive a difference when listening from a static position. Audio stimuli were generated with room acoustics modeling software. The listeners had to compare two audio samples and tell whether they perceived a difference and, additionally, which of the two samples sounded more reverberant, clearer, and more realistic. As expected, the perceptual evaluation exhibited significant individual differences when the values exceeded the JNDs. Hoare et al. (2010) assessed the perceptive discrimination ability for varying sound-source directivities in a virtual free field. Discriminative tests with direct comparison were used to evaluate the dissimilarity on a 10-point Likert scale. The listeners clearly perceived differences between the auralized sound-source directivities. Zotter et al. (2014) found that listeners are able to rotate directional sound sources towards them. Without this rotation movement, it were hardly possible to estimate the static orientation. Furthermore, the distance perception could be influenced by variations in the source directivity, since they directly affected the direct-to-reverberant energy ratio (DRR). Wendt et al. (2017) addressed this question in a virtual and a real environment.

Results of Zotter et al. (2014) let assume that listeners are able to perceive and distinguish different directivities of sound sources when being allowed to explore the area by moving past or around it. If the motion causes a change of the relative angle between the source and the listener, direction-dependent differences of the source properties may become audible. In a pilot listening test, this assumption was verified (Sloma and Neidhardt 2018) as follows.

For the assessment, sound sources with different directivities were modeled in rooms with differing room acoustics properties using MCRoomSim (Wabnitz et al. 2010). Binaural Room Impulse Responses were generated and auralized with PyBinSim (Neidhardt et al. 2017). The listeners had to state which directivity the presented sound source had, firstly when listening from a fixed position with head rotation, and secondly when listening after exploration of the room on a pre-defined walking path with head rotation allowed. The listening test was conducted in a listening laboratory according to ITU-R BS.1116. Twenty listeners participated in the experiment. The results show that listening from a fixed position did not provide sufficient information to decide whether a sound source is omnidirectional or oriented towards a specific direction. In the static listening scenario, the reverberation in the room had an influence on the decision. Walking past or around the sound source enabled the listeners to distinguish sound source directivities clearly. The relative-position changes between source and listener led to audible changes in timbre and loudness. Contrary to the static case, the room acoustic properties did not show an influence.

Further research should investigate which cues become available due to movements of the listeners, and which are actually exploited when discriminating between sound sources and their directional characteristics (Sloma and Neidhardt 2018).

Positional Disparity and the Theory of Cognitive Maps

When listening to an acoustic scene from two different static positions, the listener can benefit from the positional disparity between those two listening perspectives. For example, if the first listening position is in line with two sources and the listener has difficulties distinguishing which one is in front, a step to the side will bring a relative angular difference. In this second listening position, the listener may have fewer problems in distinguishing the sources. When walking through a scene, multiple listening perspectives add up and the sum of the information may be used for the interpretation of the acoustical cues and the understanding of the scene.

Epstein et al. (2017) assumes that humans may create a cognitive map for spatial navigation. In this way, it would become possible that interactive exploration of the acoustics is used to establish a map of the surrounding environment. However, as has been pointed out, for example, by Weisberg and Newcombe (2018), this assumption is discussed controversially. Nevertheless, Seeber and Clapp (2020) present a theory as to which listeners collect information on room geometry whenever entering a new room and build an abstract cognitive model of the room for themselves. This information helps to reduce the required adaption to room-acoustical changes while walking through the room. The adaptation to room acoustics influences, among other things, localization accuracy and speech intelligibility. In this context, it may be assumed

that extracting spatial information is easier if an exploration via interactive changes in the listening position is possible. Furthermore, it opens up further questions like

- How accurately does the adaptation of the early reflections to different positions have to be realized in a reproduction setup?
- How sensitive are listeners to simplifications in the progressive modifications of the virtual sound field while walking through a simulated environment?

4.3 Types of Interaction and Relevance of Authentic Self-motion

A person can listen passively to the motion of the sound source without having any control of this motion. In this case, no interaction is possible. A different case is given when the movements of the sound source can be controlled by the user. Then a further distinction has to be made of *authentic interaction*, where the listeners carry out an equivalent motion with parts of their bodies, for instance, tracking the source movement with their hands, and *non-authentic interaction*, where the source movement is controlled via a keyboard or equivalent simple devices.

It is of interest to consider different types of interaction with the avatar that represents the user in the virtual acoustic scene. Again, there is the option of passively listening to the moving avatar without having control or influence. If this type of interaction is used in virtual acoustic environments, it is likely that the users get confused and high concentration and listening effort are consequently required. Such cases occur in audio books or when watching videos of an avatar moving through a virtual world. Though the video at least provides some visual scene context, it will probably still be difficult to achieve the impression of immersion or to perceive the perspective of the avatar.

Furthermore, there are again the options of authentic and non-authentic interaction. Non-authentic interaction means controlling the avatar with movements and actions that are different from the motion of the avatar, such as, for instance, changing the listening position via keyboard, touchpad, or joystick. In cases of authentic interaction, the listeners control the avatar representing themselves by equivalent movements of their own body, which are tracked by a motion-capture system. For the reproduction, perceptual requirements regarding real-time interaction have then to be met, among other things, regarding latency or temporal-resolution.

Wallmeier and Wiegrebe (2014) compared three types of rotation with different degrees of interaction while the audio signal at the listeners' ears remained the same in all conditions. The participants had to rotate until being aligned as parallel as possible to a long virtual corridor. In one case, the non-moving listeners could rotate the acoustic scene around them by controlling the angular-rotation with a joystick. In the second test condition, the listeners controlled with the joystick the rotation of the chair that they were sitting in. The third condition included tracked rotation of the head in addition to the rotation of the chair. From the results of these experiments,

it may be concluded that authentic self-rotation provides additional cues compared to being moved passively. In both cases, the same audio signal is presented to the users. However, obviously, the proprioceptive information from self-motion provides additional cues that are useful for the interpretation of the auditory information received.

Genzel et al. (2018) studied the influence of a relative lateral-position change between two sound sources and the listeners in a distance-discrimination task. In one condition, the listener moved actively, in the second, the listeners were moved, and in a third, the sources moved but the listeners remained in the same position. In the case of listeners active self-motion, the smallest just-noticeable distance differences were found and for the source motion, the results were the worst. This confirms the role of self-motion also for the case of translational movements.

4.4 Effect of Exploration on Room Perception, Adaptation, and Room Divergence

The *direct-to-reverberant energy ratio* (DRR) is a relevant cue in auditory-distance estimation (Zahorik et al. 2005). However, identical DRRs in different rooms may correspond to different distances. Hence, our hearing needs to adapt to the particular room to enable accurate judgment of the distance—as discussed, for example, by Shinn-Cunningham (2000). This means that, when entering an unknown room, it is necessary to find out which DRR corresponds best to a specific distance. An exploration phase is thus likely to be beneficial for this purpose.

In conclusion, one might set up the hypothesis that an exploration phase supports the process of adaptation to a room and the abstraction of basic information about the room—compare Seeber and Clapp (2020), this volume. If this hypothesis holds, this aspect needs to be considered in connection with the room-divergence effect. It may well be that effect is strong before adapting to the new room but decreases after adaptation.

However, these thoughts are still at a hypothetical stage. Experimental evidence has not been established so far. Furthermore, the individual strategy of exploration certainly plays a significant role. Further studies are needed for clarification.

4.5 Summary of Potential Influences by Active Self-motion

In this section, potential influences of active listener translation on the interpretation of the sound-pressure signals at the ears are discussed. They can be summarized as follows.

- Interaction of self-motion and the auditory sense (effects of self-motion on the interpretation of the sound pressure at the ears)
- Benefit of positional disparity and the potential of creating a cognitive map of the environment

- Dynamic acoustical cues like current changes of intensity (acoustic τ), of the azimuth to a sound source or azimuth difference between two sources (motion parallax) or of the direct-to-reverberant ratio, may be exploited

So far, only a few studies were conducted to investigate those potential influences qualitatively and quantitatively. However, in this chapter we formulated the hypothesis that active listening may help to overcome certain deviations in localization due to non-individual HRTFs, for instance, an increased elevation angle. The listeners may perceive slight source movements during their own motion enabling them to understand and to assign the position of the source in the scene more accurately. The effect might be even stronger when the listeners change their position interactively. As yet, no studies have investigated into this question.

5 Implications for System Development

5.1 Challenges in the Realization of an Interactive Exploration of the Virtual Acoustic Scene

As regards technological application, both authentic and non-authentic interaction are challenging in different ways. For non-authentic movement, it is often difficult to achieve the perception of self-motion (*vection*; Larsson et al. 2004), especially if no visual representation of the virtual scene is provided. Authentic interaction is particularly demanding in terms of accurately capturing of the users' movements as well as keeping the system delay and temporal resolution (e.g., update rate) below the just noticeable differences—for data compare, for example, Lentz (2007). Furthermore, in technological systems, the motion capture may be restricted to certain degrees of freedom, for example, head rotation solely in the horizontal plane or position tracking only in the horizontal plane without considering height. Also, the area of action is often limited to a rather small spatial section. The restrictions are likely to affect the perceptual plausibility. However, these effects may be minimized by a system design that considers these restrictions by creating a context that communicates the reasons for the limitations and makes sense to the listeners.

Sensitivity for Auditory Motion

Perrott and Saberi (1990) determined that the *minimum-audible angle* (MAA) of the sound-source position relative to the listener is 1° in the horizontal plane and 3.6° in the vertical plane. These results are in line with observations reported by Blauert (1997) but only hold for static listeners as well as static sound sources.

In contrast, the *minimum-audible-movement angle* (MAMA) is the azimuthal displacement of a moving sound source relative to a static one that a listener can just detect (Lundbeck et al. 2017). The MAMA increases with velocity and also depends on the bandwidth and the spectrum of the acoustic signals (Carlile and Leung 2016). Results of experiments with broadband stimuli suggest that in the case of low

velocities the performance is similar to that obtained with static sources (Saberi and Perrott 1990). For angle changes in the horizontal plane, Saberi and Perrott (1990) determined MAMAs of 1.7° for a low velocity of $1.8^\circ/\text{s}$ and about 10° at for high velocity of $320^\circ/\text{s}$. For angular velocities of at least $25\text{--}100^\circ/\text{s}$, reverberation appears to have no impact.

There are no reports yet regarding the *minimum-audible-movement distance* (MAMD). However, just-noticeable differences have been determined for the *direct-to-reverberant ratio* by Zahorik (2002b), Larsen et al. (2008) and for the *sound level*, for example, by Florentine and Buus (1981).

5.2 Room-Related Binaural Synthesis

The processes and effects of auditory adaptation confirm the assumption that the construction of spatial auditory events does not depend exclusively on the synthesis technology for ear signals but also on various context-dependent quality parameters. Section 3 of the current chapter discusses auditory adaptation effects in detail. A conclusion that is drawn there is that the acoustic properties of the transfer functions used in binaural synthesis can be modified in such a way that the divergences as generated by different scene contexts can be reduced or even resolved.

Binaural Room Impulse Responses (BRIRs) can be used to reproduce sound sources in a room. The BRIRs can result from room-acoustic simulations or from measurements of real sound sources in real rooms. A comprehensive synthesis of an auditory scene with a variety of sound sources, room acoustics, and movements of the sources and receiver requires a high number of BRIRs. Minimization of the number while maintaining high perceived quality is desirable. Furthermore, it is a challenge to adjust the binaural synthesis to the prevailing listening conditions and thus to the context-dependent quality parameters.

Simplified approaches are presented which adjust single acoustic parameters of the (re-)synthesis to the acoustic parameters of the listening situation. Methods have been developed that, for example, adjust energy-based parameters, time-based parameters, or combinations of both. Driven by the goal of a perceptive fit between the perceived synthesized audio scene and its expected internal representation, the acoustic parameters of the binaural-synthesis system that are suitable for creating plausible auditory illusions are examined in the following. The objective is the development of methods for

Adjustments in accordance with the listening room The room-acoustic properties of the BRIRs measured in a room should be adjusted in accordance with the listening room. The adjustments are made by changing single or several room-acoustic parameters of the recorded BRIRs until they correspond to the listening room. The direct-to-reverberant-energy ratio (DRR) is selected as one prominent acoustic parameter for distance and room perception—compare Bronkhorst and Houtgast (1999) and Zahorik (2002a). Following the investigation on the influence

of reverberation on the externalization of auditory events (Begault and Wenzel 2001), a study is now presented that investigates the reliability of DRR adjustment for room congruence and its influence on externalization at room divergence.

Synthesis of new BRIRs New BRIRs are to be synthesized at different positions in space from measured BRIRs from other positions. The aim of the methods is to generate a large number of BRIRs from spatially sparse measurement positions and, thus, a small number of BRIRs (Brandenburg et al. 2018a). The approach focuses on the adjustment of the *initial time-delay gap* (ITDG) as the main distance cue. For other approaches to synthesize new BRIRs based on interpolations using *dynamic time warping* it is referred to the literature—for example to Sass (2012) and Pachatz (2017).

Several approaches have been made in the last years to reduce the number of measurements for the collection of data to be used for binaural synthesis. Savioja et al. (1999) describe methods to merge calculations of head-related transfer functions, acoustic room simulations, amount of reverberation, and/or source directivities. Algazi et al. (2004) developed a technique called *motion-tracked binaural* (MTB) sound for capturing, recording, and reproducing spatial sound. The authors use a circular array of microphones on a sphere with the diameter of an average head to capture the sound. When playing the sound back via headphones the movement of the listeners' head is tracked and the headphone signals are interpolated from the microphone recordings at positions close to that of the listeners' ears. This procedure, in modified form, can also be applied for interpolation between measured BRIRs. Kearney et al. (2009) simulate changes in source movement by interpolation of BRIRs. Pörschmann and Wiefing (2015), and Pörschmann et al. (2017) developed an approach that creates BRIRs by using HRTFs and omnidirectional representation of the room reverberation.

Adjustment of the Direct-to-Reverberant Energy Ratio for better externalization

Experiments are presented that investigated the reliability of DRR-adjustment to the synthesis of listening rooms in the event of convergence or divergence between the synthesized room and the listening room. The test listeners had the task of adjusting the DRR of the synthesized sound until perceptive congruence between the simulation and the individual expectation was achieved. The reference formation of the setting was based on previous training in the listening room with real loudspeakers. In addition to setting the DRR, the focus was put on the evaluation of the perceived externalization in case of room divergence after appropriate adjustment. Here are results for different settings,

- *Binaural Synthesis* The system uses measurements of individual and dummy head BRIRs for two selected rooms, sound sources, and positions. A non-dynamic system without head tracking is used to prevent dynamic cues from resolving perceptual ambiguities, such as quadrant errors and in-head localization, thus masking the effects of the DRR setting if necessary. A customizable binaural system is used to increase the fidelity of the simulation compared to real speakers. In-ear microphones are used to measure individual BRIRs and headphone transfer functions

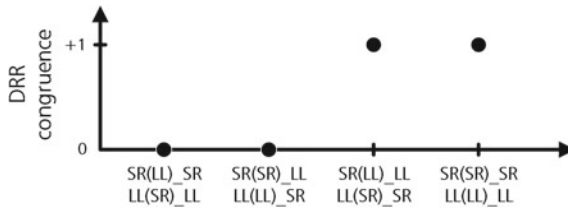


Fig. 4 Congruence of the direct sound to reverberation energy ratio (DRR) between the DRR of the adjusted synthesized room and the listening room for different room combinations. Room name in brackets denotes the DRR setting for this room; SR ...seminar room, LL ...listening laboratory; example: SR(LL)_LL is the synthesis of the seminar room (SR) in the listening lab (LL) with a DRR adaptation of the synthesis on the listening lab (LL)

(HPTFs) at the entrance of each test person's blocked ear canal. The inversion of an HPTF is calculated according to a least-square method. Extra-aural headphones (TYPE BK211; Erbes et al. 2012) are used for playback that meet the requirements for open headphones. This allows a test design with listening to the real room and resynthesis of these speakers through headphones

- *Room Divergence* A listening laboratory (abbr. LL; Rec. ITU-R BS.1116-1, $V=179\text{ m}^3$, $RT60=0.34\text{ s}$) and an empty seminar room (abbr. SR; $V=182\text{ m}^3$, $RT60=2.0\text{ s}$) were selected to take different room acoustic characteristics into account. The used combinations of listening rooms and synthesized rooms were as follows: (a), synthesis of the seminar room in the seminar room, (b), synthesis of the seminar room in the listening laboratory, (c), synthesis of the listening laboratory in the listening laboratory, and (d), synthesis of the listening laboratory in the seminar room. The adjustment of the DRRs of the synthesis by the test listeners led to further combinations of listening room and synthesized room. The used rooms were identical to those used in the experiments on the room-divergence effect reported in Sect. 2.1

Figure 4 schematically illustrates the room combinations used in the test and the congruence of the DRR between the adjusted synthesized room and the listening room. The synthesized sound sources are located at six positions around the listener. To create the binaural synthesis, the BRIRs are measured from the directions 0° , 30° , 60° , 180° , 240° , and 300° in the listening laboratory and in the seminar room. The distance of the measured speakers to the recording location is 2.2 m. The loudspeaker is directed horizontally to the recording location and is located in the median plane with an angle of 0° with respect to an assumed test listener.

Testing was performed as follows. The test listeners adjusted the DRR of the synthesis until perceptive congruence between the synthesis and the listening room was perceived. Room congruence describes the perception of auditory correspondence between the synthesized scene and the expectation on the listening room. The expectations are based on the knowledge of the listening room and are assumed as an internal reference for the comparison. The amount of room congruence was

described along the judgments *fitting accuracy*, *naturalness given*, *well sounding*, or *pleasant*. These terms are literally translated from German (Zabel 2012).

The various DRR-scaled BRIRs were changed by amplifying or attenuating the reverberation component relative to the direct-sound component of the measured individual BRIRs. The reverberation component was determined 3 ms after the direct sound. The BRIRs with different DRR-levels were calculated stepwise in a range from zero reverberation (reverberant part set to zero) up to a maximum amount of reverberation. The maximum was fixed at a reverberation level of +30% of the original measured BRIR. Seventy steps were used in between. Figure 5 shows the DRR levels as adjusted by the test listeners.

The test persons succeeded quite well in adjusting the DRR level for room convergent scenarios. The averaged deviation of the medians across all directions for the condition “LL in LL” results in a deviation of ≈ 1 dB from the measured level. The interquartile distances (IQD) averaged over all directions are at 3.8 dB. For the convergent condition (SR in SR) the median of the adjusted levels for all directions is on average 1.3 dB above the levels of measured ones. However, the adjusted DRR levels lie within the just noticeable differences (JNDs) in DRR perception (Larsen et al. 2008; Reichard and Schmidt 1966; Zahorik 2002a). The IQDs of adjustments for the condition SR in SR are on average for all directions at ≈ 5 dB. The IQDs determined in this experiment at large are ≈ 1 dB higher than the IQDs determined in a similar experiment but with experienced test listeners (Werner and Liebetrau 2014).

For the room-divergent synthesis of the listening laboratory in the seminar room (LL in SR) the median of the adjusted DRR levels across all directions 4.4 dB is on average above the DRR level of the measured ones. The test listeners’ DRR settings differ significantly from those for the resynthesis of the seminar room. The reason for this is the limited range of DRR levels used in the synthesis of the listening laboratory. The selectable DRR levels do not reach the low DRR levels as measured in the seminar room. However, the determined IQDs of the settings lie, averaged over all directions, at ≈ 2.8 dB and are therefore in the same range as for the room convergent scenarios. For the second room, for the divergent condition (SR in LL), an averaged deviation of the adjusted median over all directions from the measured DRR level of 9 dB was found. The test listeners thus chose a less reverberant synthesis than that of listening room. The IQDs averaged over all directions are at 17.5 dB. The IQD is thus significantly higher than the JNDs reported in the literature (Larsen et al. 2008; Reichard and Schmidt 1966; Zahorik 2002a). A similar test with well-experienced test persons shows much smaller IQDs of max. 8.2 dB for the same room combination (Werner and Liebetrau 2014).

The externalization results are plotted in Figs. 6 and 7 reporting the ratings of the test listeners in the quality tests for the evaluation of externalization. The results are presented as externalization indices with associated 95% confidence interval for two frontal directions and spatial combinations. The indices for the other directions show comparable behavior—not plotted.

Figure 6 shows the indices during playback in the seminar room. The highest values were achieved for the synthesis of the seminar room in the seminar room

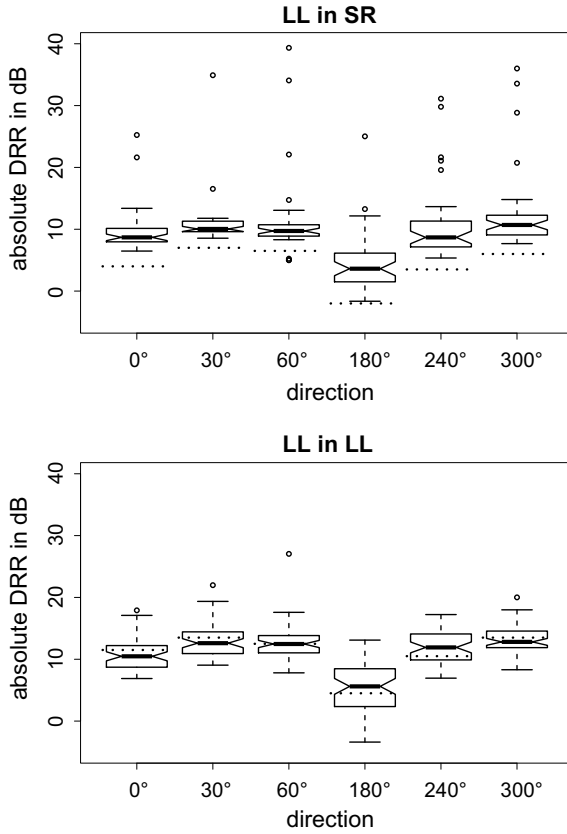


Fig. 5 Adjusted DRRs for different directions and room conditions. The **upper panel** shows the settings for the room divergent synthesis of the listening laboratory in the seminar room while the **lower one** depicts the settings for a room convergent synthesis. The adjustments for the two other room combinations show a similar behavior—not included in the figure. The DRR levels of the listening rooms are plotted as **dashed lines**

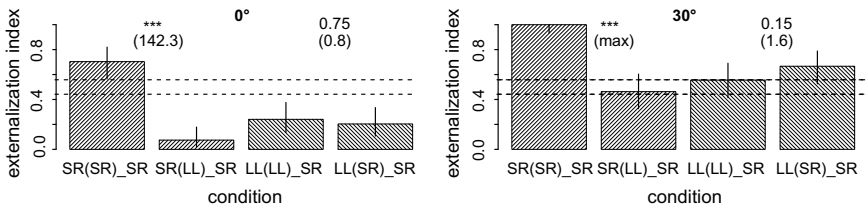


Fig. 6 Externalization indices in the seminar room with 95% confidence intervals; dashed line: 95% confidence interval of a binominal test for a selection probability of 0.5 (rate probability at N=216); SR=seminar room, LL=listening lab; *** probability value of difference at $p < 0.01$ or indication of the p-value with quota ratio

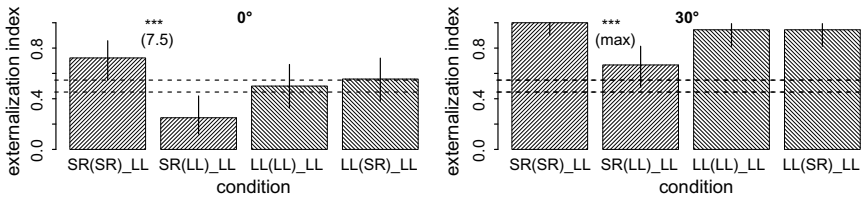


Fig. 7 Externalization indices in the listening laboratory with 95% confidence intervals; dashed line 95% confidence interval of a binominal test for a selection probability of 0.5 (rate probability at N=324); SR ...seminar room, LL ...listening laboratory, *** ...probability value of difference at $p < 0.01$ or indication of the p-value with quota ratio

and by DRR congruence (SR(SR)_SR). On average over all directions the index is about 0.9, while lower indices of about 0.7 can be found at 0° and 180°, that is, for directions with expected localization errors. A similar effect occurs for the synthesis of the seminar room in the seminar room but with DRR divergence (SR(LL)_SR). For this room condition, the indices for the 0° and 180° directions are ≈0.1. The averaged index across all directions is 0.3. For the synthesis of audio sources at lateral directions the average index is ≈0.4. For the directions 30° and 60°, the indices are in the range of the probability of guessing for the dichotomous feature externalization.

The room condition with room divergence and DRR divergence between the synthesized room and the listening room (LL(LL)_SR) reaches an index of ≈0.4 on average over all directions. The indices for the directions 30° and 300° are in the range of the rate probability. For the room condition with a divergence between the synthesized room and the listening room but DRR convergence with the listening room (LL(SR)_SR), an index of ≈0.5 is achieved on average over all directions. The indices for the directions of 0° and 180° are also lower here compared to the lateral directions, namely, ≈0.2. For the directions of 30° and 60°, the highest indices with ≈0.7 and ≈0.8 are reached. The indices for the room condition LL(SR)_SR are thus above the indices for the room condition LL(LL)_SR for all directions for which no DRR adjustment to the listening room was made.

Figure 7 shows the assessments of the test persons for the synthesis of the seminar room or the listening laboratory in the listening laboratory with DRR convergence and DRR divergence of the synthesis with the listening room. Compared to the ratings in the seminar room, higher indices are achieved for all room conditions. As expected, the lowest indices are visible for the directions of 0° and 180°. For the synthesis of the seminar room in the listening laboratory and for DRR divergence (SR(LL)_LL), the average index over all directions is ≈0.5. For the directions 0° and 180° the index drops to 0.25 and 0.2. There are no significant differences in the other room conditions. The synthesis of the listening laboratory in the listening laboratory at DRR Convergence (LL(LL)_LL) reaches high indices of over 0.9 for lateral directions.

Similar values can also be observed for the synthesis of the listening laboratory and DRR adjustment to the seminar room (LL(SR)_LL) and synthesis of the seminar room (SR(SR)_LL) with no change of the DRR. In comparison to Fig. 6 it can be seen that the synthesis of a reverberant room in a less reverberant room leads to higher indices and thus to increased externalization. The room divergence effect occurs mainly during the synthesis of a less reverberant room in a reverberant listening room.

In conclusion, the assessments of the DRR settings from the first experiment show that trained test listeners are able to reliably adjust the DRR of the synthesized listening room. The adjustment is an absolute one, made on the basis of the expected congruence of the reverberation. There is no comparison or relative alignment process between the listening room and the synthesized room. At the beginning of the test, the listeners hear the real room via loudspeakers and then, subsequently, adjust the binaurally synthesized one.

The achieved IQDs of evaluations are comparable to the JNDs at 50% detection rate found in the literature for DRR perception (Larsen et al. 2008; Reichard and Schmidt 1966; Zahorik 2002a). The IQD contains 50% of the ratings between the 1st and 3rd quantile. The low IQDs shows high reliability among the test listeners which is an indication for a suitable controllable acoustic-quality element of the synthesis. The setting of the DRR also seems to be a valid method for adjusting binaural synthesis to the reverberation of the listening room as a context-dependent quality parameter. The IQD of the DRR settings can be taken as an indication of the JND of DRR perception via the relationship $JND \leq IQA$. With the synthesized room and the listening room being in congruence, test listeners adjust a by 2 dB to 3.5 dB higher DRR for the synthesized room compared to the real listening room. The test listeners reliably choose a less reverberant synthesis. This effect is also detectable as anecdotal evidence for loudspeaker and headphone reproduction without the use of binaural synthesis.

The listening test in the second experiment investigates the effect of a DRR adjusted synthesis on the perception of externalization when there is room divergence with the listening room. The test listeners, who were not trained in this auditory test, achieved slightly larger IQDs in the DRR setting than in the first experiment. Higher DRR values (less reverberant synthesis) also tend to be set.

The ratings regarding the externalization of auditory events in case of room divergence but with the DRR of the synthesized rooms adjusted to the listening room do not provide the expected increase. Although the test stimuli adjusted to DRR-congruent signals tend to achieve higher externalization indices than those-adjusted to room-divergent ones, but these increases are generally very small with probability value of difference ($p < 0.05$). The DRR has proven to be a very reliable and adjustable room-acoustic feature in the sense of an internal reference. However, it has only a minor effect on externalization if the the room-divergence effect take effect.

Thus, it seems that the adjustment of a pure energy-based room-acoustic parameter does not lead to the desired correspondence of the stimulation patterns or schemata of the synthesis with the stored stimulation patterns or schemata and expectations. This can be explained with the Clifton effect, in which the temporal structure of

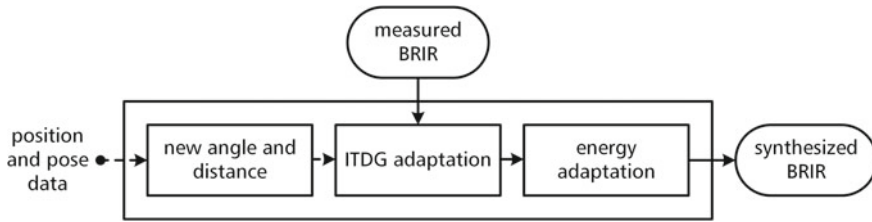


Fig. 8 Approach for synthesis of new BRIRs based on measurement of one BRIR using adjustment of the Initial Time-Delay Gap (ITDG). Figure after Mittag (2016)

the reflections is the decisive characteristic carrier. Only higher reverberation of the signals on its own does not necessarily leads to an increase of externalization.

Modifying the Initial Time-Delay Gap

The following method computes new binaural room-impulse responses (BRIRs) for desired positions in a room from a single measured BRIR dataset. The measured BRIR is adjusted to the desired new position in the room by changing the initial time-delay gap (ITDG) as one of the most prominent absolute distance cues. Figure 8 illustrates the approach. The method described is based on works of Füg (2012), Füg et al. (2012), and was further developed and evaluated by Mittag (2016), Mittag et al. (2017).

To create the new BRIRs, the position and pose data of the measured position and the new position to be synthesized are required. At the measuring position, there is a set of BRIRs for the different posing directions available. The angular resolution corresponds to the desired resolution. An interpolation of the direct sound component to increase the resolution is also conceivable. Then the distance between the position to be synthesized and the measuring position is calculated. Starting from the desired pose direction at the synthesis position, the corresponding angle at the measuring position is determined.

Based on the calculated distance change between a measured position and the position to be synthesized, the ITDG was adjusted in accordance with the measured BRIR. For this, a method was used that was developed and evaluated by Füg (2012), Füg et al. (2012), and Werner and Füg (2012). In brief, the direct sound and the first reflection were detected. An endpoint in time within the early reflections was determined by applying the perceptual mixing time. The new ITDG was applied, and the time range of the early reflections were stretched or compressed according to the new desired distance. Finally, according to the synthesized distance, the energy of the new BRIR was adjusted. Here the direct sound was adjusted according to an energy drop of 6 dB and reverberation with a drop of 1.5 dB per doubling of distance. This also achieved an implicit distance-dependent adjustment of the DRR.

Two following drawbacks of this method have to be kept in mind. First, The applied energy adjustment achieved that the energy curve over time (EDC) corresponds as far as possible to that of the measured position. The EDC was only changed by the ITDG adjustment and the associated compression or stretching of the course of the

first reflections. However, this change did not necessarily correspond to the EDC process at the synthesis position. Second, The stretching or compression of the time course of the first reflections caused a change in the spectral composition and thus a coloration compared to an assumed original BRIR. In studies by Hassager et al. (2016) it has been proven that a spectral smoothing and thus a change of the spectral composition of the reverberation has little or no influence on the externalization of auditory events.

The relevance of acoustically incorrect adaptation to spatial auditory perception can be estimated on the basis of studies regarding the identification of the listening positions in the room. Shinn-Cunningham (2003) concluded from her investigation that it was possible for trained listeners to distinguish whether they were in the middle of the room or close to a wall. Yet, a further differentiation between identification of more than one position was not easily possible.

Based on these findings, Neidhardt (2016) and Klein et al. (2017a) investigated the influence of head movements and training on the performance in position-identification tasks in a room. It was shown that targeted head movements had no significant effect on this kind of tasks. The studies on the influence of training confirm that without training the listeners cannot distinguish between different positions in a room. A training led to a significant improvement of the segregation of positions in space for 6 out of 21 (29%) listeners. However, there was also a clear dependence on the positions in the room (e.g., close to a wall or and in the middle of the room)—compare Shinn-Cunningham (2003).

Studies of Pörschmann and Wiefeling (2015) and Neidhardt et al. (2016) showed that it was difficult for listeners to correctly classify themselves on the basis of acoustic characteristics. However, it remains open to what extent, for example, in augmented audio environments, a comparison of acoustic characteristics between real sound sources in the room and synthesized virtual sound sources leads to an increased discriminatory capacity. It is also still open to what extent a real sound source, but at different positions and with different acoustic properties, can serve as a reference for virtual sound sources.

5.3 *Application Scenarios*

Binaural technology has a broad field of possible commercial applications. For example, virtualization of loudspeaker-based home entertainment systems, mobile mixing and mastering studios, spatial sound in conjecture with head-mounted displays and 360° videos, applications in wearables/hearables, gaming, and many more. In these applications, binaural sound used to be perceived as a *nice-to-have* additional feature. However, this situation has substantially changed with the advent of augmented-reality. Binaural audio processing is necessary to complement visual augmented reality or the (real) reality itself with adequate acoustics or augmented audio objects. The technical demand for this application is high. Context-dependent effects have thus to be considered.

The innovation potential of augmented acoustic reality is about to trigger relevant research, for instance, in the fields of room-acoustic estimation and modeling, Xiong et al. (2018), Kim et al. (2019), rapid HRTF acquisition or estimation, Nagel et al. (2018), He et al. (2018), source separation and audio scene classification, Cano et al. (2019), tracking as well as rendering on mobile devices and quality evaluation, Stecker et al. (2018), Cano et al. (2018), along with research on perceptual thresholds (e.g., room-discrimination thresholds Larsen et al. (2008)).

Among all well known scenarios where plausible or even authentic playback of sound via headphones is required, two examples in the field of augmented reality with regard to consumer electronics are presented in the following. Both require progress towards realistic auditory illusion as compared to the current state of the art. The next two paragraphs provide more details.

Walkable Virtual Loudspeaker Setups

The first example of an application scenario deals with an auditory augmented reality scenario in which a listener explores a scene with several virtual audio objects in a real room. Experimental setups to investigate BRIR synthesis methods for this purpose are used in the current authors' laboratory. Furthermore, the behavior of the listeners in a virtual and/or augmented-reality scene is an item of interest. The observation and the proper interpretation of this behavior, hopefully, will lead to a deeper understanding of plausibility and help to improve current quality-evaluation methods. Also, open and very relevant fields of interest are spatial mixing techniques and audio production tools for explorative sound and music installations. Beyond being used for research and test purposes, this type of setup has potential application for home use. Rendering of TV sound via headphone could become possible in a way that the listener feels like being in the real room and not having all the sound localized within the head.

Personalized Auditory Realities

Another application scenario for better binaural rendering has been introduced recently in Brandenburg et al. (2018a, b). The respective technology was termed *Personalized Auditory Reality* (PARTy) and includes an analysis of the acoustic environment, source separation, modification of audio elements, and rendering in such a way that the acoustic impression is a slight modification of the actual environment. To describe the basic idea, think of the acoustic equivalent to glasses in vision. Usually, there is a transparent rendering of the audio world around. When needed, for example, when there is too much noise around at a party, the system could separate wanted and unwanted signals, that is, for instance, people you are talking to versus background noise, and then add in, for instance, a virtual person that you are speaking to on the phone. Rendering of the signals to the two ears of the listener will be accomplished by taking advantage of all supporting conditions that have been described in the current chapter. Obviously, PARTy needs nearly perfect—at least perceptually plausible—rendering of audio sources.

6 Summary and Outlook

In this chapter experimental results have been presented that solidify the criteria needed to achieve auditory illusion over headphones. As has been known for some time, in addition to the physical reproduction of a sound field, a number of cognitive effects play a role. In fact, even when applying exactly the same sound-pressure signals to listeners' eardrums as have been presents in the recording situation, there are still a number of additional cues in the game that govern the plausibility and authenticity of the ensuing audio illusion. This includes, for example, whether the situation is known to the listeners or whether it is new, whether the listeners are experienced in listening with the corresponding head-related cues, and whether they had a chance to explore the scene by active (self-)motion.

The practical relevance of these cognitive influences has been discussed in this chapter with the aim of employing technical systems that consider the actual acoustics of the room to the end of, hopefully, delivering convincing auditory illusion.

However, research on this topic is far from being complete. While it is realized that there is a complex interaction between all the auditory cues, including, for instance, BRIRs, personalized HRTFs, reflection patterns in the actual room, learning to know the room by moving around and listening, there is still no complete model available that can reliably predict whether a desired auditory illusion will work or will break down. The modeling effort becomes even more complex since the prediction results are highly personal and, to be sure, people have widely varying individual personal experiences. In addition, the results depend on the actual audio material to be rendered.

Acknowledgements This work was partly funded by the Deutsche Forschungsgemeinschaft through the projects BR 1333/13-1 and BR 1333/18-1, as well as with fundings from the Free State of Thuringia and the European Social Fund. The authors thank two external reviewers for constructive comments and suggestions.

References

- Algazi, R., R.O. Duda, and D.M Thompson. 2004. Motion-tracked binaural sound. *Journal of the Audio Engineering Society* 52 (11): 1142–1156. www.aes.org/e-lib/browse.cfm?elib=13028. Last accessed Oct 2019.
- Ashmead, D., D. Davis, and A. Northington. 1995. Contribution of listeners' approaching motion to auditory perception. *Journal of Experimental Psychology, Human Perception and Performance* 21 (2): 239–256. <https://doi.org/10.1037/0096-1523.21.2.239>.
- Begault, D.R., and E.M. Wenzel. 2001. Direct comparison of the impact of head tracking, reverberation and individualized head-related transfer functions on the spatial perception of the virtual speech source. *Journal of the Audio Engineering Society* 49 (10). www.aes.org/e-lib/browse.cfm?elib=10175. Last accessed 10 Oct 2019.
- Bertelson, P., and M. Radeau. 1981. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception and Psychophysics* 6 (29): 578–584.

- Blauert, J. 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*, revised, 2nd ed. Cambridge, MA: MIT Press.
- Brandenburg, K., E. Cano, F. Klein, T. Köllmer, A. Lukashevich, A. Neidhardt, U. Sloma, and S. Werner. 2018a. Plausible augmentation of auditory scenes using dynamic binaural synthesis for personalized auditory realities. In *AES International Conference on Audio for Virtual and Augmented Reality*. Redmond, USA: Audio Engineering Society (AES). www.aes.org/e-lib/browse.cfm?elib=19691. Last accessed 10 Oct 2019.
- Brandenburg, K., E. Cano Ceron, F. Klein, T. Köllmer, H. Lukashevich, A. Neidhardt, J. Nowak, U. Sloma, and S. Werner. 2018b. Personalized auditory reality. In *44th Annual Meeting on Acoustics (DAGA)*. Garching by Munich, Germany: Deutsche Gesellschaft für Akustik (DEGA).
- Brimijoin, W., A. Boyd, and M. Akeroyd. 2013. The contribution of head movement to the externalization and internalization of sounds. *PLOS ONE* 8. <https://doi.org/10.1371/journal.pone.0083068>.
- Brinkmann, F., A. Lindau, and S. Weinzierl. 2017. On the authenticity of individual dynamic binaural synthesis. *The Journal of the Acoustical Society of America* 142 (4): 1784–1795. <https://doi.org/10.1121/1.5005606>.
- Bronkhorst, A., and T. Houtgast. 1999. Auditory distance perception in rooms. *Nature* 397: 517–520. <https://doi.org/10.1038/17374>.
- Bronkhorst, A.W. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica United with Acustica* 86 (1): 117–128.
- Cano, E., D. FitzGerald, A. Liutkus, M.D. Plumbley, and F. Stöter. 2019. Musical source separation: An introduction. *IEEE Signal Processing Magazine* 36 (1): 31–40. <https://doi.org/10.1109/MSP.2018.2874719>.
- Cano, E., J. Liebetrau, D. Fitzgerald, and K. Brandenburg. 2018. The dimensions of perceptual quality of sound source separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 601–605. <https://doi.org/10.1109/ICASSP.2018.8462325>.
- Carlile, S., and J. Leung. 2016. The perception of auditory motion. *Trends in Hearing* 20: 1–19. <https://doi.org/10.1177/2331216516644254>.
- Cherry, E.C. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America* 25 (5): 975–979. <https://doi.org/10.1121/1.1907229>.
- Clifton, R.K. 1987. Breakdown of echo suppression in the precedence effect. *The Journal of the Acoustical Society of America* 82 (5): 1834–1835. <https://doi.org/10.1121/1.395802>.
- Durgin, F. 2009. When walking makes perception better. *Current Directions in Psychological Science* 18 (1): 43–47. <https://doi.org/10.1111/j.1467-8721.2009.01603.x>.
- Durgin, F., et al. 2005. Self-motion perception during locomotor recalibration: More than meets the eye. *Journal of Experimental Psychology: Human Perception and Performance* 31: 398–419. <https://doi.org/10.1037/0096-1523.31.3.398>.
- Epstein, R., E. Patai, J. Julian, and H. Spiers. 2017. The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience* 20 (11). <https://doi.org/10.1038/nn.4656>.
- Erbes, V., F. Schultz, A. Lindau, and S. Weinzierl. 2012. An extraaural headphone system for optimized binaural reproduction. In *38. Jahrestagung für Akustik, DAGA*. Darmstadt, Deutschland.
- Florentine, M., and S. Buus. 1981. An excitation-pattern model for intensity discrimination. *The Journal of the Acoustical Society of America* 70: 1646–1654.
- Füg, S. 2012. Untersuchungen zur Distanzwahrnehmung von Hörereignissen bei Kopfhörerwiedergabe [Investigations on distance perception of auditory events using headphones]. Master's thesis, Technische Universität Ilmenau, Germany.
- Füg, S., S. Werner, and K. Brandenburg. 2012. Controlled auditory distance perception using binaural headphone reproduction—Algorithms and evaluation. In *27th Tonmeisterstagung, International VDT Convention*, 614–620. Cologne, Germany: Verband Deutscher Tonmeister e.V.
- Genzel, D., M. Schutte, W. Brimijoin, P.R. MacNeilage, and L. Wiegrebe. 2018. Psychophysical evidence for auditory motion parallax. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*. <https://doi.org/10.1073/pnas.1712058115>.

- Gil-Carvajal, J.C., J. Cubick, S. Santurette, and T. Dau. 2016. Spatial hearing with incongruent visual or auditory room cues. *Nature Scientific Reports* 6: 37342. <https://doi.org/10.1038/srep37342>.
- Guski, R. 1992. Acoustic tau: An easy analogue to visual tau? *Ecological Psychology* 4: 189–197. https://doi.org/10.1207/s15326969eco0403_4.
- Harris, C. 1980. Insight or out of sight? Two examples of perceptual plasticity in the human adult. In *Visual Coding and Adaptability*, 95–149. Hillsdale, NJ: Erlbaum.
- Hassager, H.G., F. Gran, and T. Dau. 2016. The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment. *The Journal of the Acoustical Society of America* 139 (5): 2992–3000. <https://doi.org/10.1121/1.4950847>.
- Hayashibe, K. 1996. The efficient range of velocities for inducing depth perception by motion parallax. *Perceptual and Motor Skills* 83: 659–674. <https://doi.org/10.2466/pms.1996.83.2.659>.
- He, J., R. Ranjan, W.-S. Gan, N. K. Chaudhary, N. D. Hai, and R. Gupta. 2018. Fast continuous measurement of HRTFs with unconstrained head movements for 3D audio. *Journal of the Audio Engineering Society* 66 (11): 884–900. www.aes.org/e-lib/browse.cfm?elib=19866. Last accessed 10 Oct 2019.
- Heeter, C. 1992. Being there: The subjective experience of presence. *Presence: Teleoperators and Virtual Environments* 1 (2): 262–271. <https://doi.org/10.1162/pres.1992.1.2.262>.
- Hendrickx, E., P. Stitt, J.-C. Messonnier, J.-M. Lyzwa, B. Katz, and C. De Boisheraud. 2017. Improvement of externalization by listener and source movement using a “binauralized” microphone array. *Journal of the Audio Engineering Society* 65 (7/8). <https://doi.org/10.17743/jaes.2017.0018>.
- Hoare, S., A. Southern, and D. Murphy. 2010. Study of the effect of source directivity on the perception of sound in a virtual free-field. In *128th AES Convention*. London, UK: Audio Engineering Society (AES). www.aes.org/e-lib/browse.cfm?elib=15369. Last accessed 10 Oct 2019.
- Hofman, P.M., J.G. Van-Riswick, and A.J.V. Opstal. 1998. Relearning sound localization with new ears. *Nature Neuroscience* 1 (5): 417–421. <https://doi.org/10.1038/1633>.
- Kearney, G., X. Liu, A. Manns, and M. Gorzel. 2015. Auditory distance perception with static and dynamic binaural rendering. In *57th International AES Conference on the Future of Audio Entertainment Technology—Cinema, Television and the Internet*. Hollywood, CA: Audio Engineering Society (AES). www.aes.org/e-lib/browse.cfm?elib=17603. Last accessed 10 Oct 2019.
- Kearney, G., C. Masterson, S. Adams, and F. Boland. 2009. Towards efficient binaural room impulse response synthesis. In *EAA Symposium on Auralization*. Espoo, Finland: European Acoustics Association (EAA).
- Keating, P., and A.J. King. 2015. Sound localization in a changing world. *Current Opinion in Neurobiology* 35: 35–43. <https://doi.org/10.1016/j.conb.2015.06.005>.
- Keen, R., and R.L. Freyman. 2009. Release and re-buildup of listeners’ models of auditory space. *The Journal of the Acoustical Society of America* 125 (5): 3243–3252. <https://doi.org/10.1121/1.3097472>.
- Kim, C., R. Mason, and T. Brookes. 2007. An investigation into head movements made when evaluating various attributes of sound. In *122nd AES Convention*. Vienna, Austria: Audio Engineering Society (AES). www.aes.org/e-lib/browse.cfm?elib=14016. Last accessed 10 Oct 2019.
- Kim, H., L. Remaggi, P. Jackson, and A. Hilton. 2019. Immersive spatial audio reproduction for VR/AR using room acoustic modelling from 360° images. In *Proceedings IEEE VR2019*.
- Klein, F., and S. Werner. 2016. Auditory adaptation to non-individual HRTF cues in binaural audio reproduction. *Journal of the Audio Engineering Society* 64 (1/2). <https://doi.org/10.17743/jaes.2015.0092>.
- Klein, F., A. Neidhardt, M. Seipel, and T. Sporer. 2017a. Training on the acoustical identification of the listening position in a virtual environment. In *143rd AES Convention*. New York, NY: Audio Engineering Society (AES). www.aes.org/e-lib/browse.cfm?elib=19231. Last accessed 10 Oct 2019.
- Klein, F., S. Werner, and T. Mayenfels. 2017b. Influences of training on externalization in binaural synthesis in situations of room divergence. *Journal of the Audio Engineering Society* 65 (3): 178–187. <https://doi.org/10.17743/jaes.2016.0072>.

- Kuhn-Rahloff, C. 2011. Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen [Processes of plausibility judgements in selected electroacoustic reproduction systems]. Ph.D. thesis, Technische Universität Berlin, Fakultät I – Geisteswissenschaften, Berlin, Deutschland. <https://doi.org/10.14279/depositonce-2790>.
- Larsen, E., N. Iyer, C.R. Lansing, and A.S. Feng. 2008. On the minimum audible difference in direct to reverberant energy ratio. *The Journal of the Acoustical Society of America* 124 (1): 450–461. <https://doi.org/10.1121/1.2936368>.
- Larsson, P., D. Västfjäll, and M. Kleiner. 2004. Perception of self-motion and presence in auditory virtual environments. In *7th Annual Workshop Presence*. Spain: Valencia.
- Lentz, T. 2007. Binaural technology for virtual reality. Ph.D. thesis, RWTH Aachen.
- Lindau, A., and S. Weinzierl. 2011. Assessing the plausibility of virtual acoustic environments. In *Forum Acusticum*, 1187–1192. Aalborg, Denmark: European Acoustic Association. <https://doi.org/10.3813/AAA.918562>.
- Litovsky, R.Y., H.S. Colburn, W.A. Yost, and S.J. Guzman. 1999. The precedence effect. *The Journal of the Acoustical Society of America* 106 (4): 1633–1654. <https://doi.org/10.1121/1.427914>.
- Lundbeck, M., G. Grimm, V. Hohmann, S. Laugesen, and T. Neher. 2017. Sensitivity to angular and radial source movements as a function of acoustic complexity in normal and impaired hearing. *Trends in Hearing* 21: 1–14. <https://doi.org/10.1177/2331216517717152>.
- Mackensen, P. 2004. Auditive localization. Head movements, an additional cue in localization. Ph.D. thesis, Technical University, Berlin, Germany.
- Macpherson, E. 2011. Head motion, spectral cues, and Wallach's 'principle of least displacement' in sound localization. In *Principles and Applications of Spatial Hearing*, 103–120. World Scientific. https://doi.org/10.1142/9789814299312_0009.
- Majdak, P., and T.W.B. Labak. 2013. Effect of long-term training on sound localization performance with spectrally warped and bandlimited head related transfer functions. *The Journal of the Acoustical Society of America* 134 (3): 2148–2159. <https://doi.org/10.1121/1.4816543>.
- Martens, W., D. Cabrera, and S. Kim. 2009. Dynamic auditory directional cues during walking: An experimental investigation using a binaural hearing instrument. In *International Workshop on the Principles and Applications of Spatial Hearing*.
- Martens, W., D. Cabrera, and S. Kim. 2011. The 'Phantom walker' Illusion: Evidence for the dominance of dynamic interaural over spectral directional cues during walking. In *Principles and Applications of Spatial Hearing*, 81–102. World Scientific. https://doi.org/10.1142/9789814299312_0008.
- Martens, W., and S. Kim. 2009. Dominance of head-motion-coupled directional cues over other cues during active localization using a binaural hearing instrument. In *10th Western Pacific Acoustics Conference*. Beijing, China.
- Martin, R., I. Witew, M. Arana, and M. Vorländer. 2007. Influence of the source orientation on the measurement of acoustic parameters. *Acta Acustica United with Acustica* 93 (3): 387–397.
- Martinson, E., and A. Schultz. 2006. Auditory evidence grids. In *International Conference on Intelligent Robots and Systems (IROS)*. Beijing, China. <https://doi.org/10.1109/IROS.2006.281843>.
- McAnally, K., and R. Martin. 2014. Sound localization with head movement: Implications for 3-D audio displays. *Frontiers in Neuroscience*. <https://doi.org/10.3389/fnins.2014.00210>.
- McGurk, H., and J. MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264: 746–748. <https://doi.org/10.1038/264746a0>.
- Mendonça, C. 2014. A review on auditory space adaptations to altered head-related cues. *Front Neuroscience* 8 (219). <https://doi.org/10.3389/fnins.2014.00219>.
- Mendonça, C., G. Campos, P. Dias, and J.A. Santos. 2013. Learning auditory space: Generalization and long-term effects. *PLoS ONE* 8 (10): e77900. <https://doi.org/10.1371/journal.pbio.0040071>.
- Mittag, C. 2016. Entwicklung und Evaluierung eines Verfahrens zur Synthese von binauralen Raumimpulsantworten basierend auf räumlich dünnbesetzten Messungen in realen Räumen [Development and evaluation of methods for the synthesis of binaural room impulse responses based on spatially sparse measurements in real rooms]. Master's thesis, Technische Universität Ilmenau, Germany.

- Mittag, C., S. Werner, and F. Klein. 2017. Development and evaluation of methods for the synthesis of binaural room impulse responses based on spatially sparse measurements in real rooms. In *43rd Annual Meeting on Acoustics (DAGA)*. Kiel, Germany: Deutsche Gesellschaft für Akustik (DEGA).
- Mourjopoulos, J. 2020. Aesthetics aspects regarding recorded binaural sounds. In *The Technology, and of Binaural Understanding*, eds. J. Blauert and J. Braasch, 455–490. Cham, Switzerland: Springer and ASA Press.
- Nagel, S., T. Kabzinski, S. Käl, C. Antweiler, and P. Jax. 2018. Acoustic head-tracking for acquisition of head-related transfer functions with unconstrained subject movement. In *Audio Engineering Society Conference: 2018, AES International Conference on Audio for Virtual and Augmented Reality*. www.aes.org/e-lib/browse.cfm?elib=19672. Last accessed 10 Oct 2019.
- Neidhardt, A. 2016. Perception of the reverberation captured in a real room, depending on position and direction. In *22nd International Congress on Acoustics (ICA)*. Buenos Aires, Argentina.
- Neidhardt, A., B. Fiedler, and T. Heinl. 2016. Auditory perception of the listening position in virtual rooms using static and dynamic binaural synthesis. In *140th AES Convention*. Paris, France: Audio Engineering Society (AES). www.aes.org/e-lib/browse.cfm?elib=18216. Last accessed 10 Oct 2019.
- Neidhardt, A., F. Klein, N. Knoop, and T. Köllmer. 2017. Flexible python tool for dynamic binaural synthesis applications. In *142nd AES Convention*. Berlin, Germany: Audio Engineering Society (AES). www.aes.org/e-lib/browse.cfm?elib=18721. Last accessed 10 Oct 2019.
- Nicol, R. 2020. Creating auditory illusions with spatial audio technologies. In *The Technology, and of Binaural Understanding*, eds. J. Braasch and J. Blauert, 581–622. Cham, Switzerland: Springer and ASA Press.
- Pachatz, N. 2017. Untersuchungen zur Relevanz raumakustischer Parameter bei Anpassung eines Binauralsynthesesystems an die Raumakustik des Abhörraumes [Investigations on the relevance of room acoustical parameters by adaptation of a binaural synthesis system to the room acoustics of the listening room]. Bachelor's thesis, Technische Universität Ilmenau, Germany.
- Parsehian, G., and B.F.G. Katz. 2012. Rapid head-related transfer function adaptation using a virtual auditory environment. *The Journal of the Acoustical Society of America* 131 (4): 2948–2957. <https://doi.org/10.1121/1.3687448>.
- Perrott, D., and K. Saberi. 1990. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America* 87: 1728–1731. <https://doi.org/10.1121/1.399421>.
- Plenge, G. 1972. Über das Problem der Im-Kopf-Lokalisation [About the problem of in-head-localization]. *Acustica* 26 (5): 241–252.
- Pörschmann, C., P. Stade, and J. Arend. 2017. Binauralization of omnidirectional room impulse responses—algorithm and technical evaluation. In *20th International Conference on Digital Audio Effects (DAFx)*, 345–352. UK.
- Pörschmann, C., and S. Wiefing. 2015. Perceptual aspects of dynamic binaural synthesis based on measured omnidirectional room impulse responses. In *International Conference on Spatial Audio (ICSA)*. Graz, Österreich.
- Raake, A., and H. Wierstorf. 2020. Binaural evaluation of sound quality and quality-of-experience. In *The Technology, and of Binaural Understanding*, eds. J. Blauert and J. Braasch, 393–434. Cham, Switzerland: Springer and ASA Press.
- Reichard, W., and W. Schmidt. 1966. Die hörbaren Stufen des Raumeindrucks bei Musik [The audible levels of spatial impression with music]. *Acustica* 17: 175–179.
- Rogers, B., and M. Graham. 1979. Motion parallax as an independent cue for depth perception. *Perception* 8: 125–134. <https://doi.org/10.1068/p080125>.
- Rosenblum, L.D., M.S. Gordon, and L. Jarquin. 2000. Echolocating distance by moving and stationary listeners. *Ecological Psychology* 12 (3): 181–206. https://doi.org/10.1207/S15326969ECO1203_1.

- Saberi, K., and D. Perrott. 1990. Minimum audible movement angles as a function of sound source trajectory. *The Journal of the Acoustical Society of America* 88 (6): 2639–2644. <https://doi.org/10.1121/1.399984>.
- Sass, R. 2012. Synthese binauraler Raumimpulsantworten [Synthesis of binaural room impulse responses]. Master's thesis, Technische Universität Ilmenau, Germany.
- Savioja, L., J. Huopaniemi, T. Lokki, and R. Väänänen. 1999. Creating interactive virtual acoustic environment. *Journal of the Audio Engineering Society* 47 (9): 675–705. www.aes.org/e-lib/browse.cfm?elib=12095. Last accessed 10 Oct 2019.
- Schymura, C., J. Rios Grajales, and D. Kolossa. 2016. Active localization of sound sources with binaural models. In *42nd Annual Meeting on Acoustics (DAGA)*. Aachen, Germany: Deutsche Gesellschaft für Akustik (DEGA).
- Seeber, B., and H. Fastl. 2004. On auditory-visual interaction in real and virtual environments. In *18th International Congress on Acoustics (ICA)*, 2293–2296. Kyoto, Japan.
- Seeber, B.U., and S. Clapp. 2020. Auditory room learning and adaptation to sound reflection. In *The Technology, and of Binaural Listening*, eds. J. Blauert and J. Braasch, 203–222. Cham, Switzerland: Springer and ASA Press.
- Shaw, B., R. McGowan, and M. Turvey. 1991. An acoustic variable specifying time-to-contact. *Ecological Psychology* 3 (3): 253–261. https://doi.org/10.1207/s15326969eco0303_4.
- Shinn-Cunningham, B. 2000. Learning reverberation: Considerations for spatial auditory displays. In *International Conference on Auditory Display*. Atlanta, Georgia, USA.
- Shinn-Cunningham, B. 2003. Identifying where you are in the room: Sensitivity to room acoustics. In *International Conference on Auditory Display*. Boston, USA.
- Simpson, W., and L. Stanton. 1973. Head movement does not facilitate perception of the distance of a source of sound. *American Journal of Psychology* 86: 151–159. <https://doi.org/10.2307/1421856>.
- Sloma, U., and A. Neidhardt. 2018. Investigations on the impact of listener movement on the perception of source directivity in virtual acoustic environments. In *44th Annual Meeting on Acoustics (DAGA)*. Garching by Munich, Germany: Deutsche Gesellschaft für Akustik (DEGA).
- Speigle, J. M., and J.M. Loomis. 1993. Auditory distance perception by translating observers, 92–99. <https://doi.org/10.1109/VRAIS.1993.378257>.
- Stecker, G.C., T.M. Moore, M. Folkerts, D. Zotkin, and R. Duraiswami. 2018. Toward objective measures of auditory co-immersion in virtual and augmented reality. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. www.aes.org/e-lib/browse.cfm?elib=19668. Last accessed 10 Oct 2019.
- Stitt, P., E. Hendrickx, and B. Katz. 2016. The role of head tracking in binaural rendering. In *29th Tonmeistertagung, International VDT Convention*. Cologne, Germany: Verband Deutscher Tonmeister e.V.
- Störig, C., and C. Pörschmann. 2013. Investigations into velocity and distance perception based on different types of moving sound sources with respect to auditory virtual environments. *Journal of Virtual Reality and Broadcasting* 10 (4). <https://doi.org/10.20385/1860-2037/10.2013.4>.
- Sutojo, S., S. Van de Par, J. Thiemann, and A. Kohlrausch. 2020. Auditory Gestalt rules and their application. In *The Technology, and of Binaural Listening*, eds. J. Blauert and J. Braasch, 33–59. Cham, Switzerland: Springer and ASA Press.
- Thurlow, W., J. Mangels, and P. Runge. 1967. Head movements during sound localization. *The Journal of the Acoustical Society of America* 42 (2): 489–493. <https://doi.org/10.1121/1.1910605>.
- Thurlow, W., and P. Runge. 1967. Effect of induced head movements on the localization of direction of sounds. *The Journal of the Acoustical Society of America* 42 (2): 480–488. <https://doi.org/10.1121/1.1910604>.
- Toole, F.E. 1970. In-head localization of acoustic images. *The Journal of the Acoustical Society of America* 48 (4): 943–949. <https://doi.org/10.1121/1.1912233>.
- Udesen, J., T. Piechowiak, and F. Gran. 2015. The effect of vision on psychoacoustic testing with headphone-based virtual sound. *Journal of the Audio Engineering Society* 63 (7/8): 552–561. <https://doi.org/10.17743/jaes.2015.0061>.

- Väljamäe, A. 2009. Auditorily-induced illusory self-motion: A review. *Brain Research Reviews* 61 (2): 240–255. <https://doi.org/10.1016/j.brainresrev.2009.07.001>.
- Väljamäe, A., P. Larsson, D. Västfjäll, and M. Kleiner. 2005. Travelling without moving: Auditory scene cues for translational self-motion. In *International Conference on Auditory Display*. Limerick, Ireland.
- Wabnitz, A., N. Epain, C. Jin, and A. Van Schaik. 2010. Room acoustics simulation for multichannel microphone arrays. In *Proceedings of the International Symposium on Room Acoustics (ISRA)*, 1–6. Melbourne, Australia.
- Wallach, H., E.B. Newman, and M.R. Rosenzweig. 1949. The precedence effect in sound localization. *The American Journal of Psychology* 62: 315–336. <https://doi.org/10.1007/s10162-014-0496-2>.
- Wallmeier, L., and L. Wiegrebe. 2014. Self-motion facilitates echo-acoustic orientation in humans. *Royal Society Open Science* 1: 140185. <https://doi.org/10.1098/rsos.140185>.
- Wang, L.M., and M.C. Vigeant. 2008. Evaluation of output from room acoustic computer modeling and auralization due to different sound source directionalities. *Applied Acoustics* 69 (12): 1281–1293. <https://doi.org/10.1016/j.apacoust.2007.09.004>.
- Weisberg, S., and N. Newcombe. 2018. Cognitive maps: Some people make them, some people struggle. *Current Directions in Psychological Science* 27 (4): 220–226. <https://doi.org/10.1177/0963721417744521>.
- Wendt, F., F. Zotter, M. Frank, and R. Höldrich. 2017. Auditory distance control using a variable-directivity loudspeaker. *Applied Sciences* 7 (7): 666. <https://doi.org/10.3390/app7070666>.
- Werner, S., and S. Fueg. 2012. Controlled auditory distance perception using binaural headphone reproduction—Evaluation via listening tests. In *27th Tonmeistertagung, International VDT Convention*, 622–629. Cologne, Germany: Verband Deutscher Tonmeister e.V.
- Werner, S., and F. Klein. 2014. Influence of context dependent quality parameters on the perception of externalization and direction of an auditory event. In *55th International AES Conference on Spatial Audio*. Helsinki, Finland: Audio Engineering Society (AES). www.aes.org/e-lib/browse.cfm?elib=17371. Last accessed 10 Oct 2019.
- Werner, S., F. Klein, and T. Harczos. 2013. Context-dependent quality parameters and perception of auditory illusions. In *4th International Symposium on Auditory and Audiological Research ISAAR*, 445–452. Denmark.
- Werner, S., F. Klein, T. Mayenfels, and K. Brandenburg. 2016. A summary on acoustic room divergence and its effect on externalization of auditory events. In *8th International Conference on Quality of Multimedia Experience (QoMEX)*. Portugal. <https://doi.org/10.1109/QoMEX.2016.7498973>.
- Werner, S., and J. Liebetrau. 2014. Adjustment of direct-to-reverberant-energy-ratio and the just-noticeable-difference. In *6th International Workshop on Quality of Multimedia Experience (QoMEX)*. Singapore. <https://doi.org/10.1109/QoMEX.2014.7138310>.
- Werner, S., and A. Siegel. 2011. Effects of binaural auralization via headphones on the perception of acoustic scenes. In *3rd International Symposium on Auditory and Audiological Research ISAAR*, 215–222. Denmark.
- Wexler, M., and J. van Boxtel. 2005. Depth perception by the active observer. *Trends in Cognitive Sciences* 9 (9): 431–438. <https://doi.org/10.1016/j.tics.2005.06.018>.
- Xiong, F., S. Goetze, B. Kollmeier, and B.T. Meyer. 2018. Exploring auditory-inspired acoustic features for room acoustic parameter estimation from monaural speech. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 26 (10): 1809–1820.
- Zabel, A. 2012. Vergleich von Hörtestmethodiken zur Beurteilung der räumlichen Wahrnehmung bei binauraler Kopfhörerwiedergabe [Comparison of listening test methods for the perception of spatial perception with binaural headphone reproduction]. Bachelor's thesis, Technische Universität Ilmenau, Germany.
- Zahorik, P. 2002a. Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America* 4 (111): 1832–1846. <https://doi.org/10.1121/1.1458027>.

- Zahorik, P. 2002b. Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America*. 112: 2110–2117. <https://doi.org/10.1121/1.1506692>.
- Zahorik, P., P. Bangayan, V. Sundareswaran, K. Wang, and C. Tam. 2006. Perceptual recalibration in human sound localization: Learning to remediate front-back reversals. *The Journal of the Acoustical Society of America* 120 (1): 343–359. <https://doi.org/10.1121/1.2208429>.
- Zahorik, P., D.S. Brunsgart, and A.W. Bronkhorst. 2005. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica United with Acustica* 91: 409–420.
- Zotter, F., M. Frank, A. Fuchs, and D. Rudrich. 2014. Preliminary study on the perception of orientation-changing directional sound sources in rooms. In *Forum Acusticum*. Kraków.

Toward Cognitive Usage of Binaural Displays



Yôiti Suzuki, Akio Honda, Yukio Iwaya, Makoto Ohuchi
and Shuichi Sakamoto

Abstract Based on acoustic input to their two ears, humans are able to collect rich spatial information. To explore their acoustic environment in more detail, they thereby move their bodies and heads to resolve ambiguities as might appear in static spatial hearing. This process is termed “*active listening*.” This chapter introduces new research regarding two specific aspects of active listening, namely, (i), facilitation of sound localization in the median plane and, (ii), augmentation of the discrimination angle for frontal auditory object. As active listening affects spatial hearing significantly, the design of systems for spatial-sound presentation requires substantial expertise in this field. In this context, a dynamic binaural display was developed that supports active listening. The display was applied to edutainment applications such as training the spatial-perception competence of visually impaired persons. Two examples were specifically investigated for this purpose, namely, a maze game and an action game. The former facilitates players’ ability to draw cognitive maps. The latter improves the sound-localization performance of players, their eye-contact frequency during conversation, and their ability to avoid approaching objects. Results suggest that binaural displays that support active listening are indeed capable of enhancing listener experience in reproduced and virtual auditory scenes.

Y. Suzuki (✉) · S. Sakamoto
Research Institute of Electrical Communication, Tohoku University,
2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan
e-mail: yoh@riec.tohoku.ac.jp

A. Honda
Faculty of Informatics, Shizuoka Institute of Science and Technology,
2200-2 Toyosawa, Fukuroi 437-8555, Japan

Y. Iwaya
Faculty of Engineering, Tohoku Gakuin University,
1-13-1 Chuo, Tagajo 985-8537, Japan

M. Ohuchi
Faculty of Comprehensive Management, Tohoku Fukushi University,
1-8-1 Kunimi, Aoba-ku, Sendai 981-8522, Japan

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_22

1 Introduction

We humans are active creatures. It is therefore quite natural for us to explore environments by moving through them to collect accurate spatial information. These inherent movements are also true for spatial hearing (Blauert 1997). The process is not restricted to audition, but may include crossmodal cues, such as from the vestibular system. The process of gathering auditory information during exploratory head and body movements is known as “*active listening*” (Suzuki et al. 2012). Numerous studies show that active listening facilitates auditory spatial perception. Wallach (1939) demonstrated that these movements provide cues for elevation-angle assessment. Also, Thurlow and Runge (1967) showed that among horizontal head turning, nodding, and pivoting, horizontal head turning is critical for proper sound-source localization. Similar findings were obtained in virtual auditory spaces. For instance, Kawaura et al. (1989, 1991) examined sound localization of a virtual sound source, spatialized at a distance of 1.5 m from the center of listener’s heads by convolving acoustic signals with head-related impulse responses (HRIRs), the time domain representation of head-related transfer functions (HRTFs). They reported that front-back and distance judgments were markedly improved when horizontal head rotations were properly reflected in the binaural signals. Indeed, numerous reports describe the facilitation of sound localization by head movement—see Perrett and Noble (1997), Iwaya et al. (2003), Toshima and Aoki (2009), Brimijoin et al. (2013). Furthermore, active listening facilitates affective cognition (Iwaya et al. 2011). The latter authors demonstrated that head rotation enhances the *sense of presence* of listeners in virtual auditory spaces.

Moreover, in the last decade, a few studies have shown that sound-image-localization accuracy is reduced by head movement. For example, Cooper et al. (2008) presented a test sound while listeners were rotating their heads. Results showed reduction of sound-localization accuracy for sound stimuli presented during head rotation, compared with that of a static condition. Leung et al. (2008) examined auditory spatial perception during rapid head motion and reported that the perceived auditory space was compressed. Honda et al. (2016) measured movement detection for a virtual sound source during listener’s horizontal head rotation. Results showed that detection thresholds were higher (i.e. worse) when listeners rotated their heads. These results urge us to further investigate the active-listening process to draw a more complete picture.

Either way, since humans must take advantage of *active listening* to appropriately understand sound environments, knowledge of active listening is relevant and indispensable for optimal and effective design of three-dimensional (3D) sound-rendering systems. In this context, the authors are particularly interested in so-called *binaural displays*. These displays are a type of 3D auditory displays that render auditory spaces by controlling the sound signals directly at the ears of the listeners. Thereby it is important to consider the listeners’ movements in order to deliver appropriate signals to the two ears. This requires taking into account the actual listener positions with respect to the sound sources. This *head-related* rendering of sound signals

is in contrast to other types of 3D auditory displays, such as wave-field synthesis (WFS) (Berkhout et al. 1993), boundary-surface control (BoSC), (Ise 1997; Enomoto and Ise 2005), and high-order Ambisonics (HOA) (Poletti 2005), for which listeners' rotational and translational movements both are naturally reflected to their ear inputs if they remain inside the listening zone.

Following the notion of head-related sound-field rendering, a middleware module for binaural display, called *Simulation Environment for 3D Audio Software* (SiFASo) (Iwaya et al. 2005, 2011) was developed in the authors' laboratory. With SiFASo, education¹ applications have been built for the purpose of training spatial-perception competence, particularly that of visually impaired people (Honda et al. 2007, 2009, 2013; Ohuchi et al. 2006).

Binaural displays are generally applicable to virtual reality applications as assistive technology. Specifically, for visually impaired persons they are of relevance for enhancing the *quality of life*—compare (Afonso et al. 2005; Iwaya et al. 2009; Picinali et al. 2014; Seki et al. 2011; Seki 2016). Experiments in which the effect of the SiFASo system was assessed reveal promising potential of binaural displays to improve spatial perception and to take advantage of some transfer effects that are useful in daily life.

2 Head Turning and Sound-Source Localization in the Median Plane

Many studies, including those described in this chapter, indicate that head movements facilitate sound localization, including front-back discrimination. One of these studies used a robot, the TeleHead (Toshima et al. 2003). “TeleHead” is an avatar robot that follows the head movements of a human in the following way. A listener in a location different from that of the robot listens to sound signals delivered from two microphones on the robot at ear positions. Listening with TeleHead improves horizontal-plane sound localization (Toshima and Aoki 2009). However, head turning improves sound localization in the median plane as well (Perrett and Noble 1997). Previous studies with the robot (Suzuki et al. 2012) revealed that it facilitates median-plane sound localization even if the rotation angles of the robot are smaller than those of human listeners.

However, it has not yet been clarified how horizontal head rotation of human listeners should be reflected to generate virtual auditory space by means of binaural displays. To clarify this issue, the effects of horizontal head rotation on sound localization were studied by the authors in greater detail. In this context the effects of horizontal rotation of the robot were investigated, whereby the turning was either in phase or in anti-phase compared to the turning of the human listener.

¹A portmanteau word composed from “education” and “entertainment”.

2.1 Experimental Procedure

Three young male adults with normal hearing participated as listeners. They were all well trained for sound-localization experiments.

The robot used was the same as that used by Suzuki et al. (2012), namely, a simplified TeleHead that follows only horizontal head turns.² This simplification was applied because horizontal head rotation (*head turning*) plays by far the most important role among the three possible rotational head movements (Thurlow and Runge 1967). A head simulator (dummy head) cast after each listeners' own head was set atop the robot. The dummy head can follow head turning up to 200%. The average system latency during operation was about 50 ms. This is slightly shorter than the detection threshold in binaural displays, and much shorter than the acceptable limit in listening tests (Yairi et al. 2007, 2008b).

Figure 1 shows a schema of the experimental setup. The TeleHead was positioned in an anechoic chamber at the center of a circular loudspeaker array set-up in the median plane. The distance between the center of the set-up and the surface of the loudspeakers was 1.5 m. The sound stimulus was a pink noise of 10 s duration including 6 ms rise and decay times using a raised cosine function, and was presented through one of the 16 loudspeakers in the median plane. The loudspeaker array was arranged with an elevation of -60° in front, climbing up to the zenith (elevation of 90°), and continuing to -60° in the rear, with 20° separation between the loudspeakers. Sound signals received at the TeleHead's two ears were reproduced by headphones (Sennheiser HDA-200) in real-time. The listeners were seated in a soundproof room next to the anechoic chamber. The sound pressure level of the stimulus was set to 70 dB for frontal sound incidence. The sound pressure levels were calibrated with an artificial ear conforming to the IEC 60318-1:2009 (B&K 4153) standard, with an adapter for circumaural earphones specified in the same IEC standard (B&K DB 0843). The force added to the adapter was set to 8 ± 1 N.

The experiment consisted of seven conditions. In one, the robot kept still, with the virtual sound source in front, that is, the static condition. In the other conditions, the robot moved in-phase or anti-phase with respect to the remote listener's horizontal head rotation. The amount of head turning of the robot with respect to the listener's head rotation was modified. The rotation ratio was selected among ± 0.05 , ± 0.1 , and ± 1.0 . Here, plus and minus signs respectively mean in-phase and anti-phase rotation in relation to that of the remote listener. Listeners were asked to move their heads freely, at least once during each trial, and to identify the loudspeaker direction from the 16 alternatives. The number of repetitions was five, and the total number of trials for each condition was 80 (16 directions \times 5 repetitions each).

²This simplified version is based on TeleHead's fourth version (Hirahara et al. 2011).

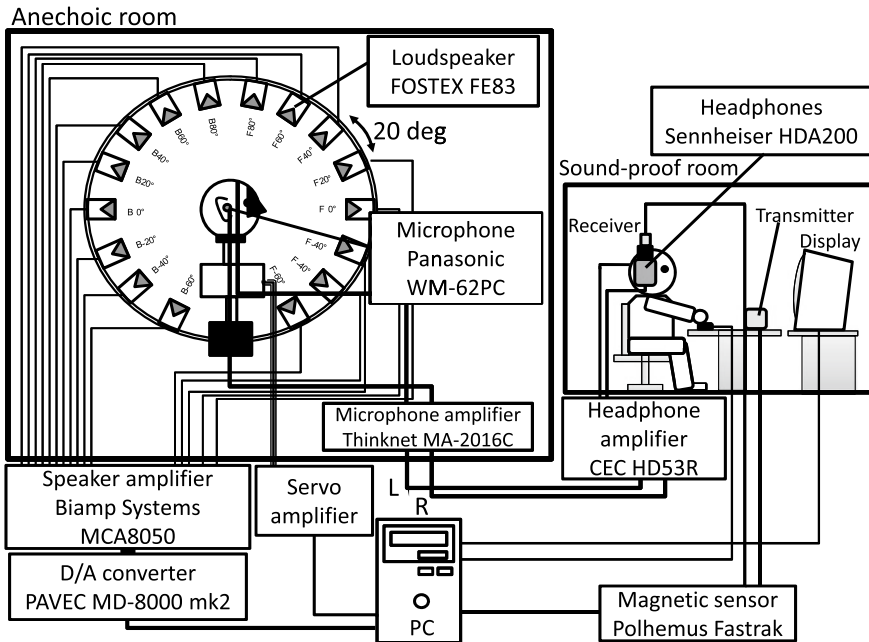


Fig. 1 Median-plane-localization setup with TeleHead

2.2 Results and Consideration

Figures 2, 3, and 4, respectively, show the localized direction as a function of physical direction in static, in-phase, and anti-phase conditions. Figure 5 shows the front-back error rate as a function of the rotation ratio for three in-phase and three anti-phase conditions compared with the static condition. Two one-way analyses of variance (ANOVA) were applied. However, Mauchly's sphericity tests indicated that both data shown in Fig. 5a, b exhibited significant departures from sphericity ($p < 0.01$). Therefore, the Greenhouse-Geisser correction factor was applied to the degrees of freedom of the ANOVA analysis. Results for in-phase rotation indicate that the effect of the rotation ratio is statistically significant ($F(1.42, 2.85) = 18.03, p < 0.05$). Multiple comparisons (Tukey's HSD, $p < 0.05$) indicate significant differences between the static and +0.1, static and +1.0, and +0.05 and +1.0 conditions. Results for anti-phase rotation indicate no significant effect ($F(1.56, 3.13) = 3.20, n.s.$).

Figure 2 shows that frequent front-back errors occurred in the static condition where TeleHead did not respond to the listener's head rotation. In contrast, Fig. 3 shows that front-back confusions were suppressed when the robot rotated in-phase to listeners' rotation, irrespective of the rotation ratio. The results of ANOVA confirm that the suppression is significant, not only when the ratio of TeleHead's rotation is 100%, but also when it is 10% of the listener's head rotation. Although the multiple comparison does not show any significant difference between the static and +0.05

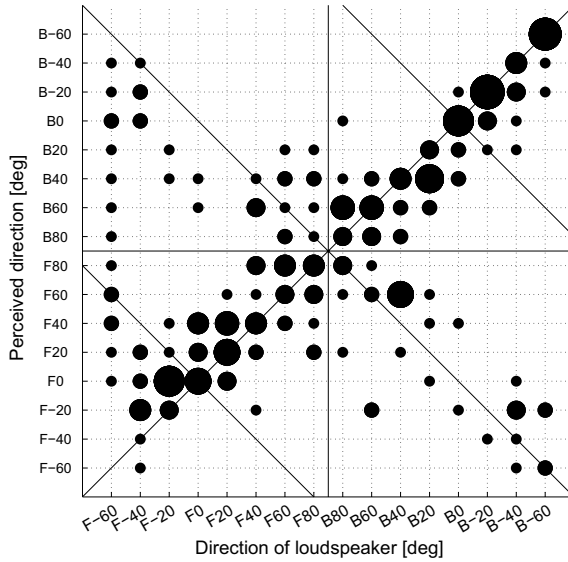


Fig. 2 Relationship between presented and localized elevation angles for the static condition ($r = 0.44$)

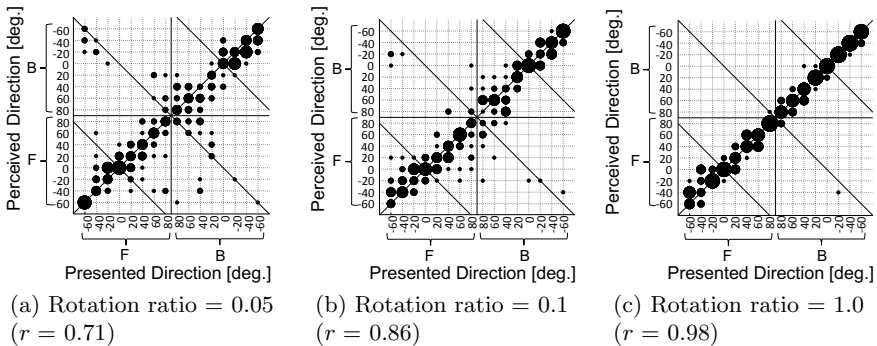


Fig. 3 Relationship between presented and localized elevation angles when TeleHead rotates in-phase

conditions, the differences in correlation coefficients between the static condition and those of the +0.05, +0.10, and +1.0 conditions are all statistically significant ($ps < 0.001$). This signifies that the in-phase feedback to the listeners results in significantly “sharper” distributions, which may imply fewer localization errors, including front-back confusions, even when the ratio is +0.05.

These results mean that an avatar robot can provide effective dynamic sound-localization cues when it rotates in-phase with the listener’s rotation, even with head turnings of the robot are as low as 5% that of the active listener.

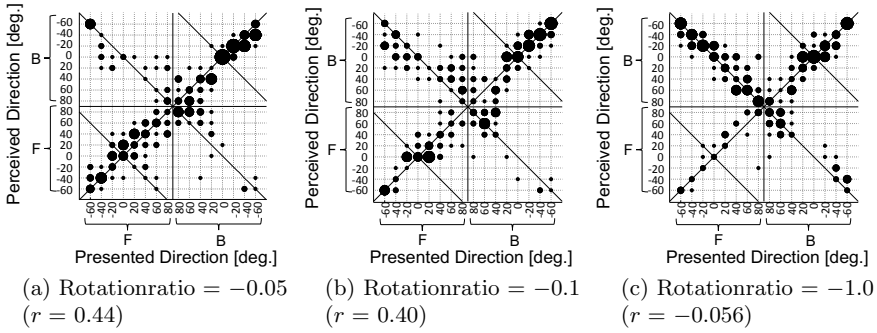


Fig. 4 Relationship between presented and localized elevation angles when TeleHead rotates anti-phase

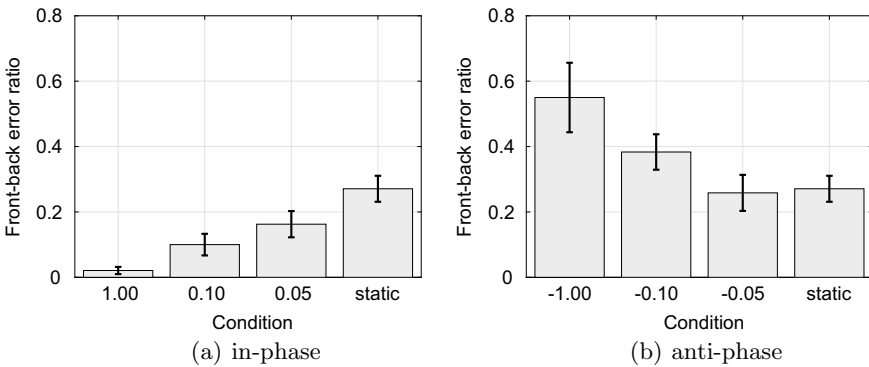


Fig. 5 Front-back error rates as a function of rotation ratio

It is noteworthy that all listeners reported that the dynamic control to reflect head rotation was unnoticeable when the ratio of the in-phase rotation was less than or equal to 10%, while they noticed the dynamic control when the ratio was 100%. Hirahara et al. (2013) showed similar results for horizontal sound localization in an experiment also using TeleHead. These results suggest that head movements can be implicitly utilized to stabilize sound localization, and that the direction of rotation is important, not just the head rotation itself.

Figure 4 shows that perceived elevation angles hardly correlate with physical sound-source direction when the rotation ratio is -1.0 (Fig. 4c). Anti-phase rotation with a ratio of 1.0 provides reversal dynamic cues in terms of front-back confusion. In fact, all three listeners reported that they often experienced such reversals, resulting in small correlation of perceived and physical directions. In contrast, localization seems hardly affected by anti-phase rotation with rotation ratios of -0.05 and -0.1 (Fig. 4a, b). This may result in robustness against unusual disturbances of static spectral cues in median-plane localization. The results of this study confirm

that horizontal rotation of the heads of listeners provides dynamic cues for proper elevation-angle localization. This suggests that this kind of rotation should be taken into account in when designing high-definition dynamic binaural displays.

3 Localization Accuracy of the Subjective-Straight-Ahead During Active Head Rotation

In most of the studies that show facilitation of sound localization by listener movements, listeners were asked to estimate over-all sound image positions after presentation of sounds and listener movements had ended. However, *localization accuracy* can be defined as a measure of the deviation of the position of a perceived auditory object from the physical position of the respective sound source. Following this notion, there have been a few studies that observed deteriorated sound-localization accuracy during listeners' head rotations when dealing with real sound sources (Cooper et al. 2008; Leung et al. 2008) as well as virtual sound sources as rendered by a binaural display (Honda et al. 2016).

There is obviously a need for more knowledge regarding sound-localization accuracy during listener motion. Thus, sound-localization accuracy regarding the *subjective-straight-ahead* was investigated in the horizontal plane precisely at the moment when listeners actively rotated their heads. The findings were compared with the static case (static condition).

3.1 Auditory Subjective-Straight-Ahead in Static Condition

In the static condition, the listeners were sitting still on a chair, but their heads were not mechanically fixed.

Experimental Procedures

Eight males and one female with normal hearing (22–40 years of age) participated. The sound stimuli consisted of 1/3-octave-noise bursts ($f_c = 1$ kHz, SPL: 65 dB when presented continuously) of 15, 30, 80, 150, and 300 ms, including 5-ms rise and decay times. An arc array of 35 loudspeakers arranged with 2.5° separation at a distance of 1.1 m from the listener (see Fig. 6) was set up. A sound stimulus was presented from one of seven loudspeakers located within $\pm 7.5^\circ$. An LED was mounted on the loudspeaker at 0° . For each trial, the LED lit first for 1 s, and then a sound stimulus followed. The experiment was conducted in an anechoic chamber which was kept dark during sessions, so that loudspeaker positions were not visible. Listeners were asked to judge whether a test stimulus was located to left or right of their subjective straight ahead exactly at the time of presentation (two-alternative forced choice). The method of constant stimuli was used, and the number of repetitions for each

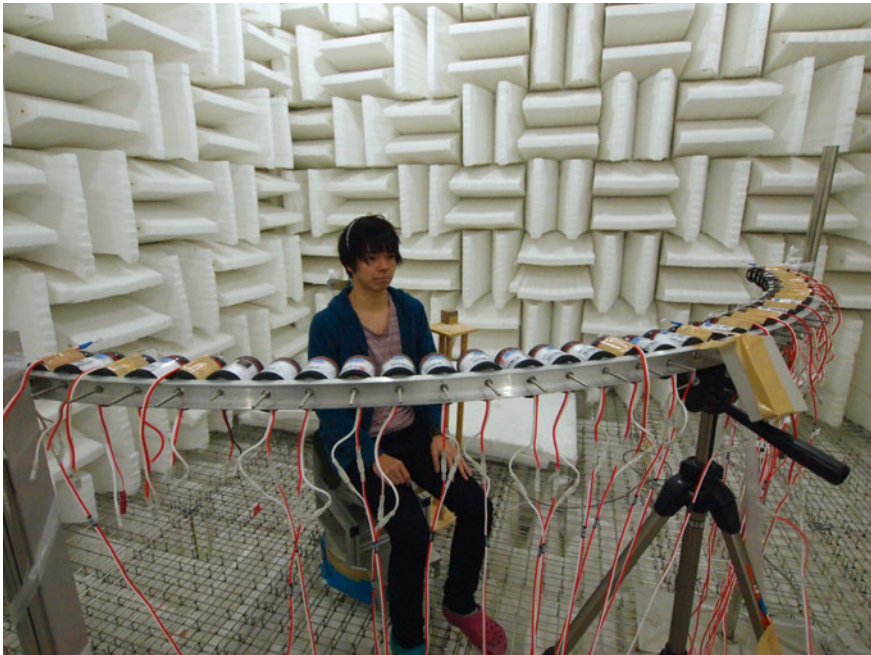


Fig. 6 Experimental setup

direction was 20. Thus, the total number of trials was 700 (5 durations \times 7 directions \times 20 repetitions).

Results

The cumulative normal distribution function was fitted to the ratios of the right-of-subjective-straight-ahead judgments as a function of stimulus direction for determining the point of subjective-straight-ahead (PSSA), and its just-noticeable difference (JND). Here, PSSA is the direction of the subjective-straight-ahead relative to the physical front.

The PSSA and its JND were respectively estimated as the mean and $0.675 \sigma^3$ of the fitted cumulative normal distribution function. Figure 7a, b respectively show PSSAs and JNDs as a function of the stimulus duration. One-way analysis of variance (ANOVA) indicates no significant difference for the direction, while the effect of the duration on JNDs are significant ($F(4, 32) = 5.29, p < 0.05$). Multiple comparisons (Tukey's HSD test, $p < 0.05$) indicate significant differences between 15 and 150 ms, and between 15 and 300 ms.

³The value of 0.675 corresponds to the z score where the cumulative normal distribution reaches 0.75, meaning an estimated correct answer rate of 75%.

3.2 Auditory Subjective-Straight-Ahead During Active Head Rotation

Sound-localization accuracy of the subjective-straight-ahead in the horizontal plane was determined with listeners sitting on a chair rotating their heads by themselves, both slowly and rapidly—that is, in active slow and active fast conditions.

Experimental Procedures

Eight of the previously tested nine listeners (seven males and one female) participated in this experiment. Examining the results of the static condition (Fig. 7), the duration of sound stimuli was set to 30 ms. This was determined because 15 ms seemed to be too short, since the JND is significantly larger for the short duration than for the longer one. However, the length should be as short as possible to minimize directional deviation during stimulus presentation. Otherwise, the experimental setup was the same as used for the static condition. Also, the method of constant stimuli was used as well. However, the method of stimulus presentation was modified to match the listeners' rotations as follows. For each trial, a guiding sound was presented from a loudspeaker located at either -45 or $+45^\circ$ for 100 ms to indicate the direction towards which the listener should rotate the head. Listeners were instructed to rotate their heads either quickly or slowly toward the direction of the guidance sound. The actual speed of rotation was observed with a motion sensors on the listeners' heads (Polhemus, Fastrak). A sound stimulus was presented when the listeners rotated their head by at least 15° . For clockwise or counterclockwise rotation, the stimulus was presented via one out of of 13 loudspeakers ranging from 0 to $+30$ or -30° . The total number of trials was 520 (2 rotational directions \times 13 stimulus directions \times 20 repetitions each).

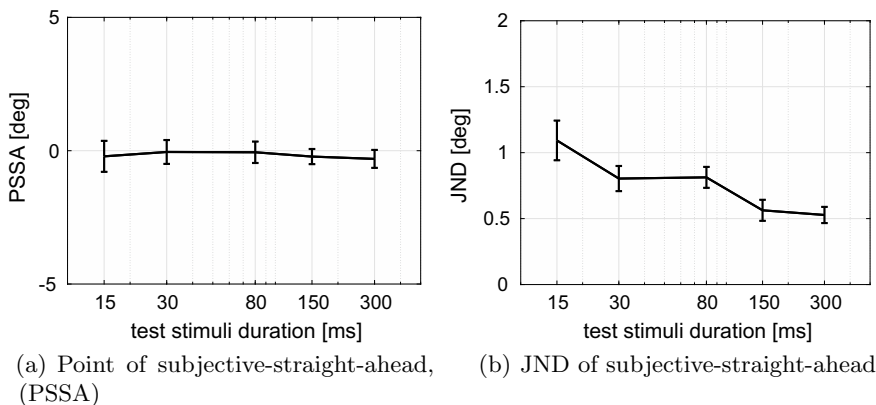


Fig. 7 Direction and JND of subjective-straight-ahead as a function of sound duration for the static condition

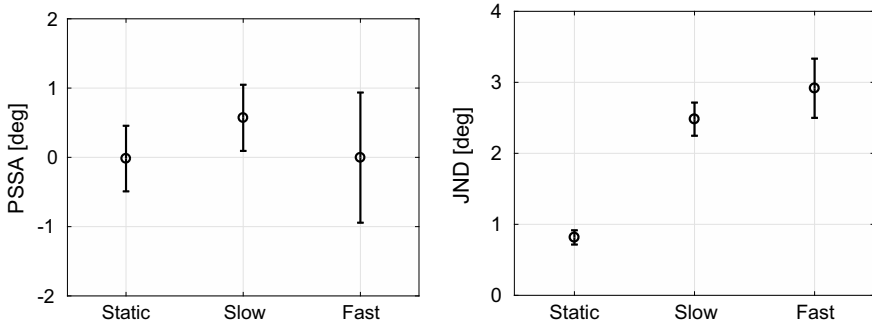


Fig. 8 Direction and JND of subjective-straight-ahead for static, active slow, and active fast conditions

Results

The means, m , and standard deviations, σ , for “slow” and “fast” head rotation speeds are $m = 12.2$, $\sigma = 4.0^\circ/s$ and $m = 166.9$, $\sigma = 69.5^\circ/s$, respectively. Paired t -test shows a significant difference between the two conditions ($t(7) = 6.07$, $p < 0.01$). This shows that the listeners could well control the two speeds. In the calculation of the PSSA and its JND, the data for clockwise and counterclockwise rotations were pooled by treating the sign of the direction toward rotation as positive, considering the symmetry of the experimental scheme.

Figure 8 depicts the PSSAs and their JNDs for the active fast and slow conditions, along with the results for the eight participants in the static condition. One-way ANOVA indicates no significance for PSSA. Alternatively, the effect of JND was significant ($F(2, 14) = 19.80$, $p < 0.01$). Multiple comparisons (Tukey’s HSD test, $p < 0.05$) indicate significant differences between the static condition and the two conditions with rotation.

3.3 Discussion

Figure 7 shows that auditory JNDs of the subjective-straight-ahead for the static condition are well below 1° when the sound duration is at least 30 ms, and are almost 0.5° when it is 150 ms or longer. This value is smaller than the minimum audible angle (MAA) in front (Mills 1958). While MAA is the difference limen of two sound images at a certain incident angle, this JND is the detection threshold for the deviation of a perceived sound object from a reference defined as directly in front. Considering this difference, the auditory subjective-straight-ahead can be regarded as very stable.

The experimental results plotted in Fig. 8b indicate that the auditory JND of the subjective-straight-ahead is significantly larger during head rotation than when lis-

teners are sitting still, irrespective of the rotation speed. Since the stimulus duration was 30 ms, the head rotated 0.36 and 5.0°, respectively, for the observed mean rotation speeds of 12 and 167°/s).

The relative drift of the direction of the sound sources, and thus that of the perceived auditory objects, can account for the increase of the JNDs of the fast condition to some extent, but this is not sufficient to also explain those of the slow condition. This may imply that the degradation is not only attributable to ambiguities of the ear-input signals induced by the movement but also to possible change of the binaural information processing in the brain for static and dynamic binaural inputs. To examine this possible difference in the mechanism of spatial hearing, further experiments for lower rotation speeds and with passive rotation should be performed.

The fact that the effects of listener's movement are not uniform is puzzling. Head motion often facilitates sound localization but may deteriorate in its accuracy, as shown here and in Sect. 2. A phenomenological explanation might be the difference in the way localization judgments occur. That is, facilitation seems to occur when overall sound localization is requested after presentation of sound stimuli and listener movements have ended, whereas deterioration is observed when instantaneous sound localization is reported during presentation while the listeners are in the course of moving. In other words, the former would resolve ambiguity in this ill-posed problem caused by scarcity of hearing inputs in only two channels (i.e. two ears), while the latter would stabilize auditory spatial perception during ear-input changes. Thus, this phenomenon can be compared to saccadic suppression in vision. Moreover, this phenomenon may be useful to design efficient 3D auditory displays, including dynamic binaural displays, because fewer computational resources can be assigned while listeners (or sound sources) are in motion.

4 A Binaural-Display Middleware, SiFASo

4.1 Binaural Displays

A binaural display is an architecture for 3D auditory display that synthesizes or reproduces the input sound signals at the listeners' ears. To realize this, sound-source signals are typically convolved with the impulse responses of the sound propagation paths from a sound source to listeners' ears. The frequency domain representation of the impulse responses of the paths can be expressed as a cascade of HRTFs and the room transfer functions (RTFs).

Psychoacoustic performance is generally good despite the simple signal processing. Moreover, as described earlier, rendering performance of binaural displays can be greatly improved by appropriately reflecting listener movements in the ear-input signals. Following the approach used in head-mounted displays (HMDs), such processing is indispensable for high-performance binaural displays to properly support active listening.

Since Morimoto and Ando (1980) first realized basic binaural displays digitally, special hardware such as digital signal processors (DSPs) had been necessary for many years to implement them (Takane et al. 1997; Blauert et al. 2000; Suzuki et al. 2002; Iwaya et al. 2002; Begault et al. 2010). However, by the 21st century it became possible to build simple binaural displays with CPUs of ordinary personal computers (Savioja et al. 1999; Wenzel et al. 2000; Miller 2001; Lokki and Järveläinen 2001; Yairi et al. 2006; Zhang and Xie 2013). In the authors' laboratory a high-performance dynamic binaural display supporting active listening as middleware, named SiFASO (Iwaya et al. 2005, 2011), has been developed. Note that so far binaural displays supporting active listening, including SiFASO, are compatible with active listening in terms of taking advantage of facilitation by listeners' movements. Designs of future dynamic binaural displays will certainly leverage current knowledge such as the suppression of sound-localization accuracy during listener movements, as discussed in Sect. 3.

4.2 *Outline of SiFASO*

SiFASO was developed based on experience with simple but low-latency (i.e. <12 ms) implementations, including the latency of position sensors (see also Yairi et al. 2006, 2008a). SiFASO can render a 3D auditory space including presentation of multiple sound sources by convolving source signals with proper individualized head-related impulse responses (HRIRs). Further, Doppler-effect (Iwaya and Suzuki 2007), 1st-order reflections, and reverberation processing are implemented. The HRIRs are interpolated to achieve smooth head and sound-source movements. Total system latency of SiFASO is about 30ms, including the head-tracker latency (Iwaya et al. 2011). Exploiting these advantages, SiFASO realizes stable, precise, and natural positioning of rendered sound images, even for moving sounds. The class diagram of the main part of SiFASO is presented in Fig. 9. SiFASO was developed as a dynamic-link library (DLL), so that it can be easily invoked from various applications. SiFASO runs under MS Windows on the CPU of a personal computer.

4.3 *Edutainment Welfare Applications for the Visually Impaired*

SiFASO was primarily developed for welfare systems to train spatial perception, particularly for visually impaired people, that is, those who must recognize spaces without having visual cues at their disposal. They are known to have better spatial hearing capabilities than sighted people. However, this sensory compensation varies with the etiology and extent of vision impairment (Paré et al. 1998), and with the age at which blindness occurs (Gougoux et al. 2004). Therefore, early support to

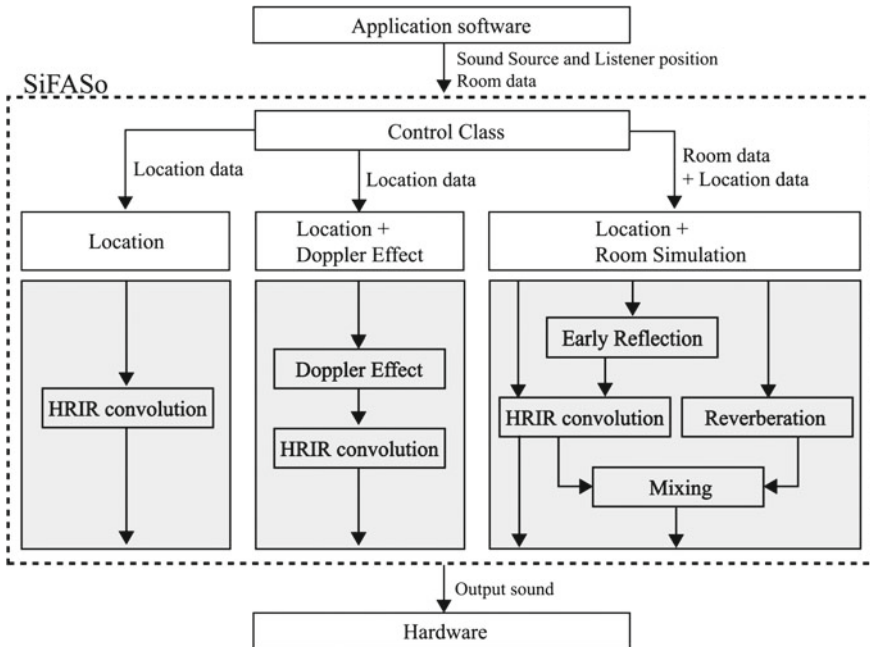


Fig. 9 Class diagram of the main part of SiFASO

improve auditory skills is extremely important, especially for people with low vision or late blindness (Afonso et al. 2005).

Three edutainment applications have been developed with SiFASO in the form of auditory virtual reality games (Iwaya et al. 2011), based on the experience of developing a similar but simpler edutainment application with dynamic binaural display using a DSP (Ohuchi et al. 2005). It is expected that they will not only be useful for training purposes but also for improving the *quality of life* of visually impaired people, who can scarcely enjoy TV games. The three edutainment applications, which are all played only by auditory information, are as follows.

- **BBBeat** An action-game-type application, where players knock out bees like a whack-a-mole game by locating their position based on a humming sound.
- **Mentalmapper** A maze-game-type application with a maze editor. Players navigate mazes rendered by spatial sounds to reach sounding landmarks assigned within the mazes.
- **SoundFormular** A racing-game-type application. Players drive vehicles and compete with a computer-controlled vehicle. Both vehicles and the motor course are rendered by spatial sounds.

All three applications were very well accepted by pupils of a municipal special-needs education school for the visually impaired. They found them to be great fun to play. Sighted pupils enjoyed SoundFormular less than BBBeat and Mentalmapper. The

courses, as rendered by SoundFormular, may have been too simple to elicit fun for those who are familiar with commercial driving games, featuring complex courses and competitors.

5 Cognitive-Map Forming in an Auditory-Maze Game

MentalMapper, an auditory-maze game (see above), was specifically developed for training and evaluating the performance of visually impaired people regarding formation of environmental cognitive maps. This section provides an outline of the game and reviews experimental results collected with it (Ohuchi et al. 2006).

5.1 Outline of the MentalMapper

The MentalMapper consists of two subsystems, namely, a maze editor and an auditory maze navigator. With the editor, mazes are drawn by connecting 1-m cube cells.⁴ For each cell, eight types of different absorption coefficients can be specified for walls, ceiling, and floor (e.g., concrete, wood, fully absorptive, and solid). Acoustic landmarks can be assigned to specified cells. These landmarks involve animal cries and environmental sounds from cars, railway crossings, etc.

With the navigator, users navigate through mazes rendered with virtual spatial sound. Navigation is performed with a game controller to move forward or backward (Fig. 10). Alternatively, users employ body rotation to turn. Further, verbal confirmations are given after each movement, such as “You have faced north.” Users hear footsteps when they move one cell forward or backward. Direct sounds and 1st-order reflections are rendered. Both auditory and tactile (vibrational) feedback are given when a user accidentally hits a wall.

5.2 Experiment 1: Evaluation of Cognitive Maps Formed via Tactile Maps

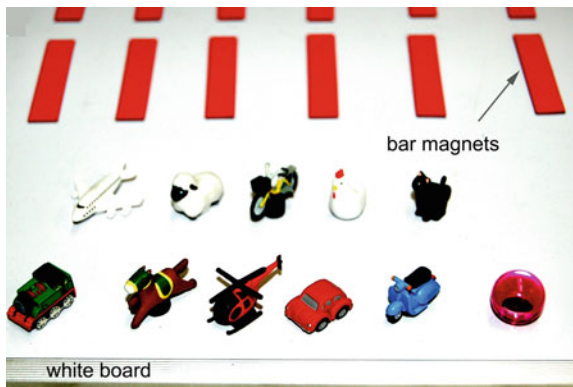
This experiment aims at examining the MentalMapper as an assistive technology for the formation of cognitive maps. Participants produce tactile maps after having navigated through virtual auditory sound mazes.

⁴This dimension was determined by technical limitations of SiFASo.

Fig. 10 Navigating a maze rendered with virtual spatial sound



Fig. 11 Whiteboard, bar magnets, and magnetic figures used in Experiment 1



Participants

Four congenitally blind adults (CMR, CSM, CTH, CKR) and four blindfolded sighted adults (SYN, SGK, SMS, SOM) participated in this experiment. Among them, one blind and all sighted participants were female.

Tactile Map

Soft and thin bar magnets were placed on a whiteboard on a desk for drawing tactile maps. Additionally, small magnetic figures (1–2 cm × 1–2 cm × 1–2 cm) were used to represent acoustic landmarks such as animals or cars in mazes (Fig. 11).

Pilot Experiment

Prior to Experiment 1, participants joined a pilot experiment. Participants were asked to navigate two auditory virtual mazes and then locate the landmarks on a blank

Fig. 12 An example of the mazes used in the pilot experiment. The cell with a star indicates the start position, and cells with circles denote landmarks

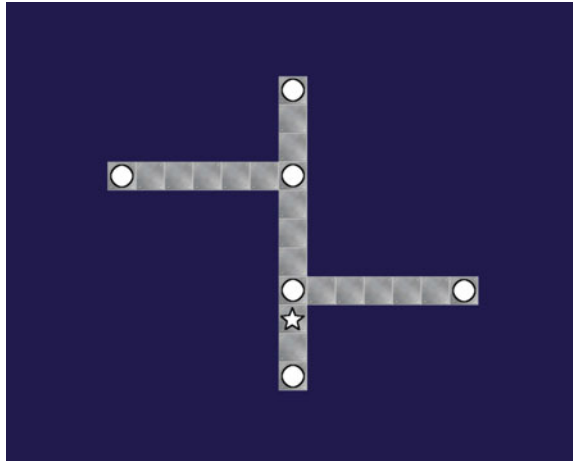


Fig. 13 Tactile-map drawing by a participant



tactile map. An example maze and a scene where a participant was drawing its map are respectively shown in Figs. 12 and 13. This was the very first experience for all of participants of moving through virtual auditory spaces by means of a game controller. While one sighted participant confused the locations of two landmarks, the others made no mistakes at all. Some of the congenitally blind participants reported that they had difficulties in creating a mental spatial image of the route, probably due to a lack of experience with such tasks.

Tasks

In Experiment 1, participants were first asked to navigate virtual auditory mazes with several landmarks and to then draw the map of the auditory maze as a tactile map. Figure 14 shows two mazes used in the task. The participants freely traversed the virtual auditory maze back and forth and then drew the tactile map. The time for the task was unrestricted. The landmark sounds were set to audible only when the user

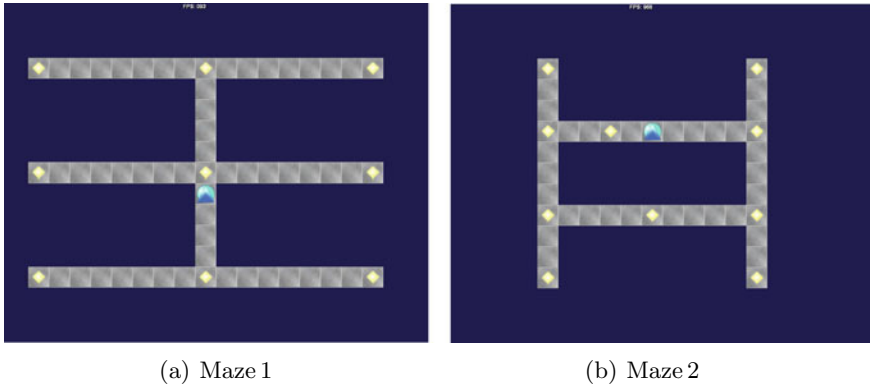


Fig. 14 Two mazes used in Experiment 1

entered the respective cell. The absorption coefficient of concrete was used for all cell boundaries.

Results and Consideration

The times required to complete the task are shown in Table 1. The blind participants spent 9' 53" and 24' 03" on average to complete Mazes 1 and 2, respectively, while the blindfolded sighted participants spent 15' 33" and 28' 01" on average. A two-way ANOVA was performed on the times to complete the task considering the group (blind or blindfolded sighted) and the maze (Maze 1 or 2) as factors. The results indicate no significant differences between blind and sighted participants ($F(1, 6) = 0.53$), although the blind group spent less time on average than the blindfolded sighted group for both mazes.

Moreover, both groups took significantly longer time on Maze 2 than on Maze 1 ($F(1, 6) = 11.61$, $p < 0.05$). This is probably due to the redundant structure of Route 2, namely that the passage of this maze causes some confusion because it forms a square walking path. Figure 15 shows examples of the maps drawn by congenitally blind and blindfolded sighted participants. All except CTH and SYN drew geometrically accurate maps. CTH and SYN could not draw complete maze shapes, but the geometry of the drawn parts was accurate. The drawn maps were then evaluated quantitatively by calculating bi-dimensional correlation coefficients between the shapes of the virtual mazes and digitized shapes drawn by participants (Tobler 1977). For the incomplete mazes of CTH and SYN, only the completed parts were analyzed. Table 2 shows the correlation coefficient for each participant on each route. All participants showed correlation coefficients greater than 0.85, suggesting that virtual auditory navigation of mazes is indeed effective in assisting formation of cognitive maps.

Table 1 Time required to complete task of Experiment 1

	Congenitally blind					Blindfolded sighted				
	CMR	CSM	CTH	CKR	Av.	SYN	SGK	SMS	SCM	Av.
Maze 1	7' 32"	15' 37"	9' 50"	6' 32"	9' 53"	18' 23"	20' 55"	10' 55"	11' 58"	15' 33"
Maze 2	12' 27"	45' 34"	24' 40"	13' 32"	24' 03"	23' 59"	47' 31"	25' 51"	14' 43"	28' 01"

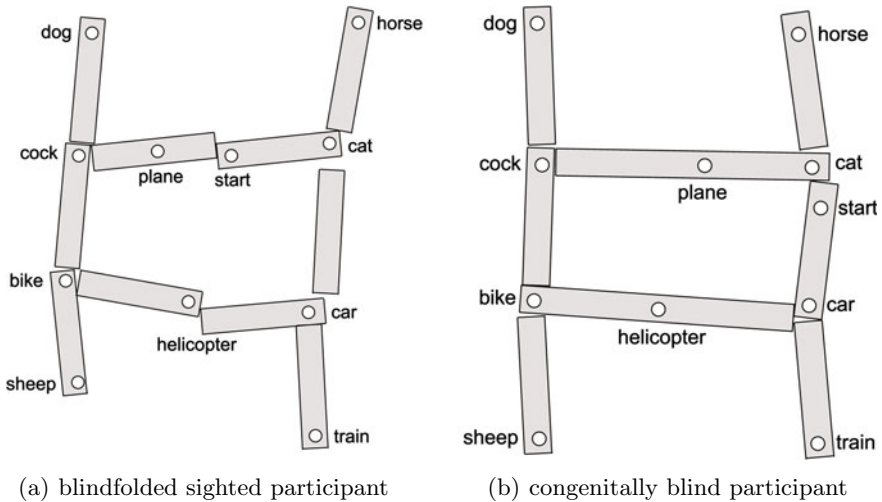


Fig. 15 Example of the maps drawn by two participants (Maze 2)

Table 2 Bidimensional correlation coefficient for each participant in each maze

	Congenitally blind				Blindfolded sighted			
	CMR	CSM	CTH	CKR	SYN	SGK	SMS	SCM
Maze 1	0.861	0.948	0.927	0.997	0.994	0.990	0.982	0.981
Maze 2	0.944	0.983	0.947	0.995	0.966	0.985	0.974	0.967

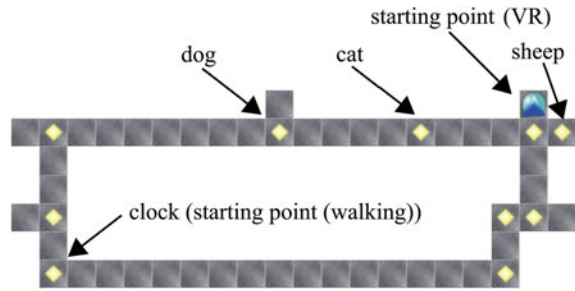
5.3 Experiment 2: Forming a Cognitive Map of an Actual Building

The results of Experiment 1 indicate that cognitive maps seem to be shaped correctly after navigation through auditory virtual environments. Experiment 2 examined whether training with auditory virtual maps is beneficial for navigating the real world.

Tasks

The maze used in this experiment was a replica of the corridor structure of an actual university building (Fig. 16). This experiment used the same four congenitally blind

Fig. 16 Visual representation of an auditory maze based on the floor structure of an actual corridor



adult subjects of Experiment 1. The sound assigned to landmarks was made audible only when the user was virtually present in the respective cell.

Participants first navigated freely through the auditory virtual maze up to 20 min to explore the maze. Then, immediately following navigation, they were asked to walk through the actual maze (i.e. a real corridor structure). To start, they were guided to the position of the landmark clock, which is different from the start position for the virtual navigation (Fig. 16). Then, the following four sequential tasks were assigned. During the tasks, for safety reasons, participants walked with a sighted guide and were asked to indicate vocally whenever changing their heading.

- **Task 2-1** Go to the landmark “dog.”
- **Task 2-2** Go to the landmark “sheep” without encountering the cat.
- **Task 2-3** Go to the landmark “dog,” taking the shortest route.
- **Task 2-4** Go back to the landmark “clock,” taking the shortest route.

Results and Discussion

This experiment aimed at examining whether participants were able to generate cognitive maps after navigating the virtual auditory maze. Task 2-2 requires participants to develop a good mental representation of the floor plan, including landscape locations. Therefore, if all these tasks are successfully completed, the cognitive maps can be regarded as well formed.

Table 3 shows the evaluation of Task 2-2 walking tasks by the experimenter for each participant. The rating scale is defined as follows:

- 4 Participants found the most direct route to the destination without wandering astray
- 3 Participants wandered but finally reached the destination
- 2 Participants reached the destination with some verbal assistance after wandering for some time or going off course
- 1 Participants could not find a way to the destination.

CSM, in Tasks 2-1 and 2-2, and CKR, in Task 2-1, wandered around before reaching the destinations. CTH became disoriented in Task 2-2. After asking for verbal assistance, CTH reached the destination and completed Tasks 2-3 and 2-4. CMR did not reach the correct destinations in Tasks 2-1 through 2-3. In Task 2-4, CMR reached

Table 3 Evaluation of Experiment 2

		Task			
		2-1	2-2	2-3	2-4
Participant	CSM	3	3	4	3
	CTH	4	2	4	4
	CKR	3	4	4	4
	CMR	1	1	1	2

the clock via an incorrect route. After the experiment, CSM and CKR were asked about their reasons for wandering back and forth in Task 2-1. CSM reported being confused because the starting point of the real environment (i.e. the clock) was not provided for the free navigation in the virtual environment. CKR reported that he walked from one end of the corridor to the other one to find out the map scale, which was not divulged to the participants.

Why could CMR not complete these experimental tasks after showing good performance in Experiment 1? A salient difference between these two experiments is the tactile maps. In Experiment 1, participants drew tactile maps after navigating through the mazes. This procedure might have reinforced the formation of cognitive mapping. If so, it suggests that the combination of navigation of virtual spaces and map-drawing is very effective in forming suitable and robust cognitive maps. Another reason could be differences in walking style. After the experiment, CMR reported difficulty in walking with a sighted guide, since she was used to walking with a guide dog in daily life. CMR was the only person having such a dog. This suggests a possible influence of the use of a guide dog on the acuity of spatial cognition in orientation and mobility, including the formation of cognitive maps. This observation raises interesting questions for future research.

Other participants mentioned that they became disoriented by physically turning right and left repeatedly in the virtual environment. This disorientation might be attributable to mental rotation in incomplete maps under formation. Typically, the physical experience while walking is key to the formation of cognitive maps (Herman et al. 1982). In contrast, repeated rotation without any real physical walking might have induced such confusion. Exploring optimal procedures for the formation of cognitive maps of virtual auditory environments is certainly a further interesting area for future research.

Overall, the experimental results show that blind participants are able to form dependable cognitive maps via virtual navigation of unfamiliar environments. Furthermore, the results show that the experience to navigate virtual auditory mazes can transfer to the ability of navigating real environments with similar geometries. These results mean that dynamic 3D auditory displays, including binaural ones, are an effective assistive tool to improve orientation and mobility of visually-impaired people by adequate training.

In this context, Seki et al. (2011) made important contributions. Among other things, they provide a training system with more realistic virtual worlds, namely, with authentic traffic sounds. These are important cues that give blind people better mobility. This system, the *Wide-Range Auditory Orientation Training System*, (WR-AOTS), is described in Seki (2016).

6 Transfer Effects from Playing the Auditory Action Game

As mentioned in Sect. 4, BBBeat is an action-game edutainment software for training sound-localization skills. In this section, transfer effects⁵ as a result of playing this auditory game are described and discussed (Honda et al. 2007, 2009).

6.1 Transfer Effects and Spatial-Hearing Training

Transfer effects are commonly observed for various motor- or verbal-learning tasks. For instance, previous studies reported transfer effects of playing visual-action video games (Castel et al. 2005; Fery and Ponsérre 2001; Green and Bavelier 2003). However, few studies have examined transfer effects of playing auditory-action games. BBBeat is an auditory action-game-type edutainment application resembling the “whack-a-mole.” The players virtually hear the hum of honeybees instead of seeing annoying moles. They are then prompted to localize the honeybee position and to hit it with a hammer as quickly and accurately as possible (Fig. 17). It has been observed that players move their heads frequently to detect the position of the hum. When hitting a bee, vibration feedback is given, and another honeybee is spawned. Honda et al. examined the various transfer effects from playing BBBeat using pre- and post-test performance results of blindfolded individuals. In the experiments, participants were separated into two groups, maintaining the same proportion of males to females. Participants of the training group were asked to play the game for seven days (30 min per day) within a two-week period. In contrast, the control group did not play the game at all during this period.

Based on the results of this experiment, transfer effect with regard to sound-localization performance for real sound sources were examined (Honda et al. 2007), and a follow-up test was conducted to investigate the persistence of the transfer effects. The task was to identify a sound source among 36 loudspeakers distributed around the listener. Results revealed that the hit rate of the training group increased by approximately 20%, which is around twice that of the control group (statistically significant). Interestingly, a follow-up test, which was conducted one month later, showed that transfer effects persisted.

⁵**Transfer effect** This is defined as the ability to extend what has been learned in one context to new contexts. This is also called “transfer of learning.”

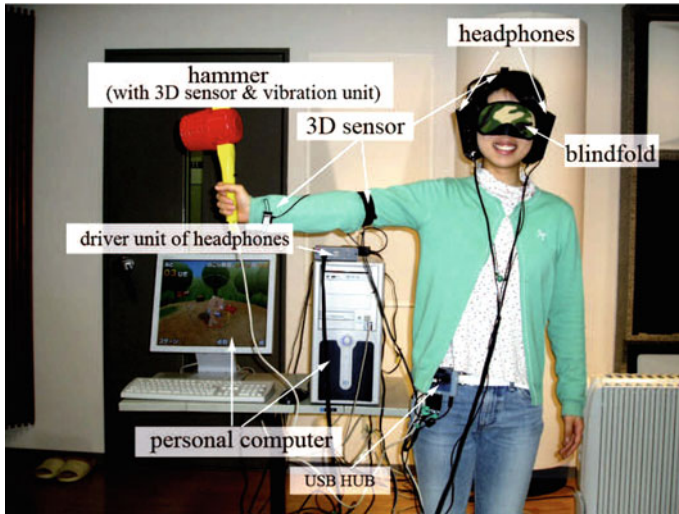


Fig. 17 Participant putting on attachments to play BBBeat

Readers may think that the above results are quite intuitive, since sound-localization performance was improved by the previous sound-localization training with BBBeat. However, the findings indicate a further interesting aspect, that the results show clear effects in the real world even though they were cultivated by training in a virtual environment. It must be admitted, however, that this transfer effect occurred between very similar tasks. Therefore, even better examples of transfer effects regarding skills that are useful in daily life are described in the following subsections (Honda et al. 2009).

6.2 Transfer Effects on Face Contacts

Normally sighted people devote attention to nonverbal information in interpersonal communication. For example, eye contact in face-to-face situations plays a regulatory function in everyday conversation (Kendon 1967). Eye contact is a relevant critical component of rewarding social exchange for sighted people (Ellsworth and Ludwig 1972). In contrast, visually impaired people use more non-visual cues for social interaction (Fichten et al. 1991). The difference in communication cues affects the impressions of visually impaired people.

Several researchers attempted to find effective training methods for the communication skills of visually impaired people (Erin et al. 1991; Sanders and Goldberg 1977; Raver 1987). For example, Sanders and Goldberg (1977) proposed a training program using auditory feedback for correct eye/face contacts to increase the rate of eye contact. These findings suggest that the communication skills could be enhanced

by training sound-localization skills. They reported that the eye contacts of clients, consisting of almost totally blind men, increased to over 80%, and the effect remained at the 74% level after 10 months.

These findings suggest that communication skills can be improved by training sound-localization skills. Therefore, Honda et al. (2009) examined whether sound-localization training is transferred to the rate of face contacts.

Experimental Conditions

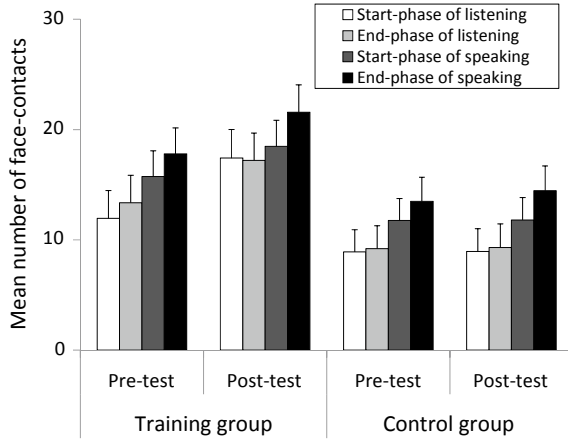
In this experiment, all blindfolded participants conducted the task in a soundproof room. Participants sat on a chair allowing free head and body movement. They were asked to discuss some topics with two interviewers (pre-test). Two trained female experimenters assisted as interviewers to ensure uniformity of tasks. Two video cameras recorded the scenes from the interviewers' positions. The interviewers provided scripted questions for the participants. Each interviewer was asked to confirm whether the participants showed face-contact behaviors on each topic with four phases, namely, (i) the start-phase of listening, (ii) the end-phase of listening, (iii) the start-phase of speaking, and (iv) the end-phase of speaking. Another interviewer confirmed whether the participants showed face-contact behavior to the querying interviewer during question-and-answer communication. Reliability obtained using the corresponding rate between the interviewers was 83%. All participants were asked to perform the same task again two weeks later (post-test). In the post-test, several topics were altered and the position of the interviewers was exchanged.

Results and Discussion

Figure 18 shows the results of the experiment. A three-way ANOVA was performed on the number of face-contacts in the communication task, considering the group (training or control), the test phase (pre-test or post-test), and the interview phase (start phase of listening, end phase of listening, start phase of speaking, or end phase of speaking) as factors. Results indicate that interaction between the group and the test phase is significant ($F(1, 37) = 5.71, p < 0.05$). The interaction can be observed in the results shown in Fig. 18. Post-hoc analysis (Ryan method, $p < 0.05$) reveals that the face-contact of the training condition ($m = 14.71$) increased significantly after playing the bee-hitting virtual auditory game ($m = 18.67, p < 0.01$). Additionally, the training group in the post-test showed more face-contact than the control group in the post-test ($m = 11.13, p < 0.05$). However, interaction between the group and the interview phase and interaction between the test and the interview phase are not significant. Furthermore, no significant three-way interaction was found.

These results indicate that, by playing the bee-hitting virtual auditory game rendered by a dynamic binaural display, face-contacts in social interaction increased significantly. This indicates that skills acquired while playing the auditory virtual game transferred to participants' communication skills during social interaction.

Fig. 18 Number of face contacts in the communication task for the training group in comparison to those of the control group



6.3 Transfer Effects on Collision-Avoidance Behavior

In order to avoid looming objects, the generation of an appropriate response includes five tasks, that is, (i), detection of a looming stimulus, (ii), localization of the stimulus position, (iii), computation of the direction of the stimulus movement, (iv), determination of an escape direction and, (v), selection of proper motor actions (Liaw and Arbib 1993). Previous studies revealed that visual information tends to be used more efficiently than auditory information regarding accuracy of estimating time-to-arrival (Schiff and Oldak 1990). However, visually impaired people are obviously restricted in their use of visual cues. Consequently, it is very important for them to formulate and execute avoidance behaviors using acoustical information for the correct location of approaching objects (i.e. perceived sound sources). Furthermore, when an object is on a collision course toward persons, they must move aside with minimal distance from their own position to that of the object, because avoidance with greater distances might cause another collision with surrounding obstacles. Therefore, when visually impaired people try to conduct appropriate avoidance behaviors, it is crucial that they perceive sound-source positions accurately. Appropriate avoidance behavior thus relies on good sound-localization skills. Consequently, Honda et al. (2009) was interested in whether sound-localization training with virtual auditory games are transferred to avoidance behaviors in response to approaching auditory objects.

Experimental Conditions

Figure 19 illustrates the collision-avoidance task. In this scene, all blindfolded participants were asked to avoid an approaching object when they felt that it was moving on a collision course (relevant path). Furthermore, they were asked to perform avoidance maneuver with minimal displacement from their position. They were further instructed not to avoid an approaching object when they felt it was moving on an irrelevant path. The distance between the relevant path and the two irrelevant paths was 80 cm. The colliding object was a toy car (width: 30 cm, weight: 2.5 kg).

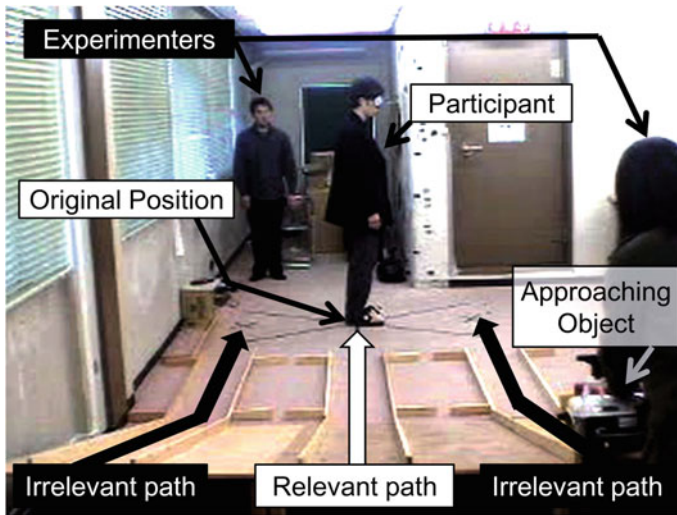


Fig. 19 Illustration of a collision-avoidance task

The approaching stimulus was presented randomly and repeatedly to participants from either relevant or irrelevant paths. The A-weighted sound pressure level of the approaching sound was 75 dB, and the background noise level in the room was 35 dB. The toy car was placed at 50 cm height on lanes (initial velocity: 0 m/s) and slid along the lane slopes at 2 m/s. Three lanes were used. The center lane was used for the relevant (i.e. collision) path, and lanes of both sides were for irrelevant paths. The distance between the participants and the lanes was 4.0 m. The task of the participants was to localize the approaching object solely based on auditory cues and to decide their behaviors within 2 s. The trials numbered 36 in all. The approaching object was sent on one of the three lanes, selected randomly, 12 times for each course. The body direction of the participants was changed for each trial. Consequently, the toy car approached from either front, back, left, or right. The experimenter then checked whether participants had completed the avoidance behaviors for each trial. Additionally, the distances from the participants' start position to the end point of their actions was measured. All participants were asked to perform the same task again two weeks later (post-test).

Results and Discussion

Figure 20 shows the results of the experiment. A three-way ANOVA was performed on the mean avoidance distances from the original position for the object approaching from irrelevant paths considering the group (training or control), the test phase (pre-test or post-test), and the direction to the approaching object (front, back, left, or right) as factors. Results show that a two-way (group \times test) interaction is significant ($F(1, 25) = 6.93, p < 0.05$). Post hoc analysis (Ryan method, $p < 0.05$) reveals

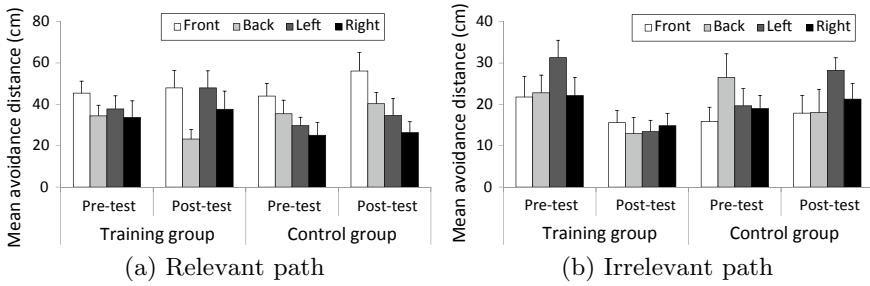


Fig. 20 Mean avoidance distances in the avoidance task

that avoidance distances of the training group ($m = 24.51$) decreased significantly by playing the bee-hitting virtual auditory game, BBBeat ($m = 14.19, p < 0.01$).

The results indicate that the avoidance distance after sound-localization training in the virtual auditory game decreased for objects approaching from the irrelevant path.

These findings indicate that auditory training using the bee-hitting virtual auditory game with binaural display modified the detailed manners of executed avoidance behaviors, which relate to sound-localization skills.

7 Concluding Remarks

In this chapter some interesting aspects of human active listening have been described. The question raised is, how can active listening be defined when considering these aspects? Listener movements induce dynamic ear inputs. As an operational definition, the following is proposed:

Active listening is a mode of multisensory spatial hearing that takes advantage of dynamic information induced by listeners' movements, irrespective of being intentional, conscious, or unconscious.

In fact, as shown in Sect. 2, even unconscious dynamic change of ear inputs may significantly change listener experience during spatial hearing.

While evidence has accumulated confirming that active listening facilitates sound-localization performances, a few recent studies have revealed that it may suppress sound-localization accuracies (see Sects. 1 and 3). However, these examples should not be regarded as an inconsistency, but rather an indication of the diversity of the roles of active listening.

Facilitation of sound-localization performance seems to occur when overall sound localization is requested after presentation of sound stimuli, and the listeners have terminated their direction-finding movements, whereas deterioration is observed when instantaneous sound-image positions are reported while listeners are moving during the sound presentation. In other words, the former would resolve ambiguities in the ill-posed problem caused by the scarcity of acoustic input to only two channels (i.e. the two ears), while the latter helps stabilize the perceptual auditory space during variation of the ear-inputs signals. This is in a way similar to *saccadic suppression* in vision. Knowledge of active listening and the psychoacoustic effects going with it are key to advancing binaural technologies. For instance, knowledge of how sound localization in dynamic scenarios can be enhanced can be directly applied to performance enhancement of, for example, dynamic 3D auditory displays, including binaural ones. In this context, knowledge of suppression and masking effects supports the economic use of computational resources during listener movements.

In Sects. 5 and 6, two edutainment applications of dynamic binaural displays were introduced. These are applied to support active listening for the training of auditory spatial-perception acuity, particularly of visually impaired people. One of the applications is a maze game and the other is an action game. The maze game facilitates users' ability to draw cognitive maps, as well as the evaluation of this capability. Moreover, a transfer effect was found to the navigation of real environments having a similar geometry as the virtually experienced maze. Playing the action game improved players' sound localization performances. Again, the experience at virtual sound localization in playing the action game transferred to improve players' sound-localization performances of real sound sources. Moreover, clear transfer effects to skills useful in daily life were observed, including increased eye-contact frequency during conversation and improved ability to avoid an approaching object. These results indicate good potential for application of dynamic binaural displays to improve spatial-hearing abilities and, hopefully, other skills that have the potential to enhance quality of life.

In summary, active listening plays an important role in making human spatial hearing more reliable and richer. Binaural technologies that support active listening are key to high-definition communication. Dynamic binaural displays that support active listening are universally applicable to enhance quality of experience in virtual and real auditory dynamic scenes.

Acknowledgements Results introduced in this chapter have been partly supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) Grant-in-Aid for Specially Promoted Research (19001004), the Society for the Promotion of Science (JSPS) *Kakenhi* Grant-in-Aid for Scientific Research (A) (24240016, 16H01736), and the Consortium R & D Projects for Regional Revitalization of the Ministry of Economy Trade and Industry from 2003 to 2004. The authors would like to thank M. Cohen for his diligent proofreading of earlier versions of the manuscript, which greatly improved the readability of this chapter. Thanks are further due to two anonymous reviewers for their very constructive comments and advice.

References

- Afonso, A., B.F. Katz, A. Blum, C. Jacquemin, and M. Denis. 2005. A study of spatial cognition in an immersive virtual audio environment: Comparing blind and blindfolded individuals. In *Proceedings of International Conference on Auditory Display*.
- Begault, D.R., E.M. Wenzel, M. Godfroy, J.D. Miller, and M.R. Anderson. 2010. Applying spatial audio to human interfaces: 25 years of NASA experience. In *40th International Audio Engineering Society Conference*.
- Berkhout, A.J., D. de Vries, and P. Vogel. 1993. Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America* 93: 2764–2778.
- Blauert, J. 1997. *Spatial Hearing*. Cambridge: The MIT Press. ISBN 0-262-02413-6.
- Blauert, J., H. Lehnert, J. Sahrhage, and H. Strauss. 2000. An interactive virtual-environment generator for psychoacoustic research. I: Architecture and implementation. *Acta Acustica United with Acustica* 86 (1): 94–102.
- Brimijoin, W.O., A.W. Boyd, and M.A. Akeroyd. 2013. The contribution of head movement to the externalization and internalization of sounds. *PLoS ONE* 8: e83068.
- Castel, A.D., J. Pratt, and E. Drummond. 2005. The effects of action video game experience on the time course of inhibition of return and the efficiency of visual search. *Acta Psychologica* 119 (2): 217–230.
- Cooper, J., S. Carlile, and D. Alasis. 2008. Distortions of auditory space during rapid head turns. *Experimental Brain Research* 191: 209–219.
- Ellsworth, P.C., and L.M. Ludwig. 1972. Visual behavior in social interaction. *Journal of Communication* 22 (4): 375–403.
- Enomoto, S., and S. Ise. 2005. A proposal of the directional speaker system based on the boundary surface control principle. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)* 88: 1–9.
- Erin, J.N., K. Dignan, and P.A. Brown. 1991. Are social skills teachable? A review of the literature. *Journal of Visual Impairment and Blindness* 85: 58–61.
- Fery, Y.-A., and S. Ponsérre. 2001. Enhancing the control of force in putting by video game training. *Acta Psychologica* 44 (12): 1025–1037.
- Fichten, C.S., D. Judd, V. Tagalakakis, and R. Amsel. 1991. Communication cues used by people with and without visual impairments in daily conversations and dating. *Journal of Visual Impairment and Blindness* 85: 371–378.
- Gougoux, F., F. Lepore, M. Lassonde, P. Voss, R.J. Zatorre, and P. Belin. 2004. Neuropsychology: Pitch discrimination in the early blind. *Nature* 430: 309.
- Green, S.C., and D. Bavelier. 2003. Action video game modifies visual selective attention. *Nature* 423: 534–537.
- Herman, J.F., R.G. Kolker, and M.L. Shaw. 1982. Effects of motor activity on children's intentional and incidental memory for spatial locations. *Child Development* 53 (1): 239–244.
- Hirahara, T., Y. Sawada, D. Morikawa. 2011. Impact of dynamic binaural signals on three-dimensional sound reproduction. In *Proceedings of Inter-Noise 2011*.
- Hirahara, T., D. Yoshisaki, and D. Morikawa. 2013. Impact of dynamic binaural signal associated with listener's voluntary movement in auditory spatial perception. In *Proceedings of Meetings on Acoustics*, Vol. 19.
- Honda, A., H. Shibata, J. Gyoba, K. Saito, Y. Iwaya, and Y. Suzuki. 2007. Transfer effects on sound localization performances from playing a virtual three-dimensional auditory game. *Applied Acoustics* 68: 885–896.
- Honda, A., H. Shibata, J. Gyoba, Y. Iwaya, and Y. Suzuki. 2009. Transfer effects on communication and collision avoidance behavior from playing a three-dimensional auditory game based on a virtual auditory display. *Applied Acoustics* 70: 868–874.
- Honda, A., H. Shibata, S. Hidaka, J. Gyoba, Y. Iwaya, and Y. Suzuki. 2013. Effects of head movement and proprioceptive feedback in training of sound localization. *i-Perception* 4: 253–264.

- Honda, A., K. Ohba, Y. Iwaya, and Y. Suzuki. 2016. Detection of sound image movement during horizontal head rotation. *iPerception* 7: 2041669516669614.
- Ise, S. 1997. A principle of active control of sound based on the Kirchhoff-Helmholtz integral equation and the inverse system theory. *Journal of the Acoustical Society of Japan* 53: 706–713.
- Iwaya, Y., and Y. Suzuki. 2007. Rendering moving sound with the doppler effect in sound space. *Applied Acoustics* 68 (8): 916–922.
- Iwaya, Y., Y. Suzuki, and D. Kimura. 2002. The effects of head movement on sound localization with real and virtual sound sources. In *Proceedings of China-Japan Joint Conference on Acoustics (JCA2002)*.
- Iwaya, Y., Y. Suzuki, and D. Kimura. 2003. Effects of head movement on front-back error in sound localization. *Acoustical Science and Technology* 24: 322–324.
- Iwaya, Y., M. Toyoda, and Y. Suzuki. 2005. A new rendering method of moving sound with the doppler effect.
- Iwaya, Y., M. Otani, and Y. Suzuki. 2009. Development of virtual auditory display software responsive to head movement and a consideration on deration of spatialized ambient sound to improve realism of perceived sound space. In *Proceedings of International Workshop on Principles and Applications of Spatial Hearing (IWPASH)*.
- Iwaya, Y., M. Otani, and Y. Suzuki. 2011. Development of virtual auditory display software responsive to head movement and a consideration on deration of spatialized ambient sound to improve realism of perceived sound space. In *Principles and Applications of Spatial Hearing*, ed. Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and H. Kato. Singapore: World Scientific.
- Kawaura, J., Y. Suzuki, F. Asano, and T. Sone. 1989. Sound localization in headphone reproduction by simulating transfer function from the sound source to the external ear (in Japanese). *Journal of the Acoustical Society of Japan* 45: 755–766.
- Kawaura, J., Y. Suzuki, F. Asano, and T. Sone. 1991. Sound localization in headphone reproduction by simulating transfer function from the sound source to the external ear (English translation). *Journal of the Acoustical Society of Japan (E)* 12: 203–216.
- Kendon, A. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26: 22–63.
- Leung, J., D. Alais, and S. Carlile. 2008. Compression of auditory space during rapid head turns. *Proceedings of the National Academy of Sciences of the United States of America* 105: 6492–6497.
- Liaw, J.-S., and M.A. Arbib. 1993. Neural mechanisms underlying direction-selective avoidance behavior. *Adaptive Behavior* 1 (3): 227–261.
- Lokki, T., and H. Järveläinen. 2001. Subjective evaluation of auralization of physics-based room acoustics modeling.
- Miller, J.D. 2001. Slab: A software-based real-time virtual acoustic environment rendering system. In *Proceedings of International Conference on Auditory Display*.
- Mills, W.A. 1958. On the minimum audible angle. *Journal of the Acoustical Society of America* 30: 237–246.
- Morimoto, M., and Y. Ando. 1980. On the simulation of sound localization. *Journal of the Acoustical Society of Japan (e)* 1 (3): 167–174.
- Ohuchi, M., Y. Iwaya, Y. Suzuki, and T. MuneKata. 2005. Training effect of a virtual auditory game on sound localization ability of the visually impaired.
- Ohuchi, M., Y. Iwaya, Y. Suzuki, and T. MuneKata. 2006. Cognitive-map forming of the blind in virtual sound environment In *Proceedings of International Conference on Auditory Display*.
- Paré, N.L.M., F. Lepore, and M. Lassonde. 1998. Early-blind human subjects localize sound sources better than sighted subjects. *Nature* 395: 278–280.
- Perrett, S., and W. Noble. 1997. The contribution of head motion cues to localization of low-pass noise. *Perception & Psychophysics* 59: 1018–1026.
- Picinali, L., A. Afonso, M. Denis, and B.F. Katz. 2014. Exploration of architectural spaces by the blind using virtual auditory reality for the construction of spatial knowledge. *International Journal of Human-Computer Studies* 72: 393–407.

- Poletti, M.A. 2005. Three-dimensional surround sound systems based on spherical harmonics. *Journal of the Audio Engineering Society* 53: 1004–1025.
- Raver, S.A. 1987. Training blind children to employ appropriate gaze direction and sitting behavior during conversation. *Education and Treatment of Children* 10 (3): 237–246.
- Sanders, R.M., and S.G. Goldberg. 1977. Eye contacts: Increasing their rate in social interactions. *Journal of Visual Impairment and Blindness* 71: 265–267.
- Savioja, L., J. Huopaniemi, T. Lokki, and R. Väänänen. 1999. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society* 47: 675–705.
- Schiff, W., and R. Oldak. 1990. Accuracy of judging time to arrival: Effects of modality, trajectory, and gender. *Journal of Experimental Psychology: Human Perception and Performance* 16 (2): 303–316.
- Seki, Y. 2016. Wide-range auditory orientation training system. <https://staff.aist.go.jp/yoshikazu-seki/AOTS/WR-AOTS/index.html> (last accessed December 21, 2019).
- Seki, Y., Y. Iwaya, T. Chiba, S. Yairi, M. Otani, M. Ohuchi, T. Munekata, K. Mitobe, and A. Honda. 2011. Auditory orientation training system developed for blind people using PC-based wide-range 3-d sound technology. In *Principles and Applications of Spatial Hearing*, ed. by Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and H. Kato. Singapore: World Scientific. ISBN: 978-981-4465-41-0.
- Suzuki, Y., S. Takane, S. Takahashi, and T. Miyajima. 2002. A preliminary development of high definition virtual acoustic display based on advise.
- Suzuki, Y., T. Okamoto, J. Treviño, Z. Cui, Y. Iwaya, S. Sakamoto, and M. Otani. 2012. 3d spatial sound systems compatible with human's active listening to realize rich high-level Kansei information. *Interdisciplinary Information Sciences* 18: 71–82.
- Takane, S., T. Miyajima, Y. Yamada, D. Arai, Y. Suzuki, and T. Sone. 1997. An auditory display based on virtual sphere model. In *Proceedings of International Symposium on Simulation Visualization and Auralization for Acoustic Research and Education*.
- Thurlow, W.R., and P.S. Runge. 1967. Effect of induced head movements on localization of direction of sound. *Journal of the Acoustical Society of America* 42: 480–488.
- Tobler, W. 1977. Bidimensional regression: A computer program. https://www.geog.ucsb.edu/~tobler/publications/pdf_docs/Bidimensional-Regression.pdf (last accessed December, 2019).
- Toshima, I., and S. Aoki. 2009. Sound localization during head movement using an acoustical telepresence robot: Telehead. *Advanced Robotics* 23: 289–304.
- Toshima, I., H. Uematsu, and T. Hirahara. 2003. A steerable dummy head that tracks three-dimensional head movement: Telehead. *Acoustical Science and Technology* 24: 327–329.
- Wallach, H. 1939. On sound localization. *Journal of the Acoustical Society of America* 10: 270–274.
- Wenzel, E.M., J.D. Miller, and J.S. Abel. 2000. Sound lab: A real-time, software-based system for the study of spatial hearing. In *Proceedings of AES 108th Convention*, p. Preprint:5140.
- Yairi, S., Y. Iwaya, and Y. Suzuki. 2006. Investigation of system latency detection threshold of virtual auditory display. In *Proc. of ICAD 2006—12th Meeting of the International Conference on Auditory Display*, 217–222, London.
- Yairi, S., Y. Iwaya, and Y. Suzuki. 2007. Estimation of detection threshold of system latency of virtual auditory display. *Applied Acoustics* 68: 851–863.
- Yairi, S., Y. Iwaya, and Y. Suzuki. 2008a. Individualization feature of head-related transfer functions based on subjective evaluation. In *Proc. of International Conference on Auditory Display (ICAD2008)*, Paris. 2008.
- Yairi, S., Y. Iwaya, and Y. Suzuki. 2008b. Influence of large system latency of virtual auditory display on behavior of head movement in sound localization task. *Acta Acustica united with Acustica* 94: 1016–1023.
- Zhang, C., and B. Xie. 2013. Platform for dynamic virtual auditory environment real-time rendering system. *Chinese Science Bulletin* 58 (3): 316–327.

Audition as a Trigger of Head Movements



Benjamin Cohen-Lhyver, Sylvain Argentieri and Bruno Gas

Abstract In multimodal realistic environments, audition and vision are the prominent two sensory modalities that work together to provide humans with a best possible perceptual understanding of the environment. Yet, when designing artificial binaural systems, this collaboration is often not honored. Instead, substantial effort is made to construct best performing purely auditory-scene-analysis systems, sometimes with goals and ambitions that reach beyond human capabilities. It is often not considered that, what enables us to perform so well in complex environments, is the ability of: (i) using more than one source of information, for instance, visual in addition to auditory one and, (ii) making assumptions about the objects to be perceived on the basis of a priori knowledge. In fact, the human capability of inferring information from one modality to another one helps substantially to efficiently analyze the complex environments that humans face everyday. Along this line of thinking, this chapter addresses the effects of *attention reorientation* triggered by audition. Accordingly, it discusses mechanisms that lead to appropriate motor reactions, such as head movements for putting our visual sensors toward an audiovisual object of interest. After presenting some of the neuronal foundations of multimodal integration and motor reactions linked to auditory-visual perception, some ideas and issues from the field of a robotics are tackled. This is accomplished by referring to computational modeling. Thereby some biological bases are discussed as underlie active multimodal perception, and it is demonstrated how these can be taken into account when designing artificial agents endowed with human-like perception.

B. Cohen-Lhyver (✉) · S. Argentieri · B. Gas
CNRS, Institut des Systèmes Intelligents et de Robotique,
ISIR, Sorbonne Université, 75005 Paris, France
e-mail: cohen.lhyver@gmail.com

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_23

1 Introduction

Assume the following situation: a listener in a lecture hall attends a talk of a fellow researcher. The conference room is almost full, and everyone has reached a seat. Yet, people keep sparingly moving in all along the talks, trying to make as little noise as possible while they thread their way through the rows to find an available chair. While the talk is still going on, the sound of a small glass breaking on the floor of the lecture hall reaches the listener from the right. The sound is sharp and vivid, but muffled due to the distance. A first observable reaction, will very likely to be the *turn-to reflex*, namely, listeners quickly turn their heads towards the object that has caused the sound.

The reason for this reflex is that such head movements are an attempt to guide the optical sensors (eyes) to spatial areas of interest, namely, to enable an analysis complementary to the one that has already been performed beforehand by the auditory modality. This primary analysis is indeed responsible for the alerting mechanism. Reactions triggered in such a way are generally termed *attention reorienting*. In the case discussed here, the reaction was initiated by auditory cues—see Fig. 1. Turning our head in a case like this is a manifestation of the need to *focus* on a particular object of interest that occurs in an environment.

Attention reorienting is an observable consequence of the integration of multiple complex mechanisms giving humans the ability to react quickly to complex environments. In particular, head movements are triggered by various signals and situations, for instance, by danger signals, but also by unexpected perceptual objects such as stimuli requiring our attention or carrying an interest with respect to a task to accomplish.

In the situation described above—besides the notion of danger signal—the main characteristics of the “falling glass” object is its obvious rareness in the context given and, consequently, its low predictability. In other words, neither any perceptual clue



Fig. 1 *Attention reorientation* caused by the occurrence of an unpredictable stimulus leading to head movement towards the audiovisual source. This motor reaction enables the visual sensors to acquire supplemental data about the object of interest—after Corbetta et al. (2008)

nor any prior information has supported anticipation of this event. However, whatever the origins of the head movements are, they resulted in putting the optical sensors towards an area that required more in-depth analyses of the perceptual information.

It has to be asked: “*What does actually ‘cause’ head-turn reactions?*” As stated above, audition is a modality capable of triggering movements of the head towards an unexpected object. However, will that reaction also occur in a situation were glasses are falling constantly and breaking on the ground? In other words, how important is the context in which an object is occurring, and how does this context matters for relevant consequent motor reactions? The same object can thus either trigger motor reactions in a specific environment or remain utterly unnoticed in a different environment. A dog barking in a kennel would undoubtedly provoke a different reaction in a persona than if barking in a person’s bedroom, given that it is not the person’s own dog. The difference between these two environments is the *predictability* of the object to occur. Thus, it has to be concluded that the occurrence of an auditory, visual, or audiovisual object is not an inherent attribute of the corresponding signals. It is the context that determines consequent motor reactions. Predictability is thus a key in understanding the mechanism of attention reorienting.

In particular, all these considerations are of importance when it comes to the design of artificial agents endowed with human-like multimodal perception capabilities. Such agents aim at understanding complex environments similarly as humans do. Thus, they have to be able to process the different kinds of signals as perceived by their dedicated sensors, artificial ears, and eyes for instance, but also to know how to combine them appropriately to form a multimodal perceptual world. The technologies of both sound processing and image processing, that is, a *multimodal approach* are needed to provide these robots with adequate comprehension of the world. However, at least in the robot community, audition and vision are often considered as two separate senses with distinct information channels, each used to form perceptual worlds in their particular way.

In order to provide more evidence of the relevance of a thorough multimodal understanding of the world, the question will be addressed of “*How can audition be utilized as a trigger for head movements towards objects of interest?*” that is, how can one modality, for example audition, be used for requisition of another modality, for instance vision, to the end of gaining a better understanding of a multimodal environment? Three key neuronal phenomena are being discussed in Sect. 2 to address this question. They all form together a solid basis of the comprehension of multimodal integration, motor reactions, and prediction abilities of the human sensory cortical areas. Understanding these mechanisms provide helpful hints for designing artificial intelligence aimed at being integrated into robots that are to be furnished with human-like perception. As an example, a computational model of the head-turn reflex driven by auditory information, called the Head-Turning Modulation (HTM) model, will be described in Sect. 3. A short conclusion ends this chapter.

2 Neuroscientific Foundations

There is extensive literature available concerning the relevant phenomena mentioned above. It includes binaural audition and sound processing by dedicated cortical areas, binocular vision, and image processing by other dedicated cortical areas, multimodal integration, attention computing, and motor reactions—both in reflexive as well as in reflective behavior—compare Blauert and Brown (2020), this volume.

Consequently, the following descriptions are restricted to biological foundations of *attention reorientation caused by audition*. Four neuronal mechanisms are dealt with in this context. These are mechanisms that represent primary biological components to be understood and considered when designing artificial agents with attentional capabilities driven by multimodal perception.

2.1 Superior Colliculus

The *Superior Colliculus* (SC), is an excellent example for illustrating how vital multimodal integration is in the brain's analysis of sensory information. It is now widely accepted that multimodal integration is crucial even for unimodal perceptual flow analysis (Atilgan et al. 2018), in particular when it comes to designing artificial systems that use auditory, visual, or any other sensory modality. Taking, for example, the *cocktail party effect* (Cherry 1953; Cherry and Taylor 1954) and the detailed analysis of auditory and visual information, the following two approaches to cross-modal interaction are conceivable.

- One may consider auditory and vision as two distinct modalities being processed separately through different and well-characterized channels, and only the results of these analyses in each perceptual modality being used for further analyses and integration
- One may, alternatively, consider that cross-modal integration is already performed at low levels of the participating modal pathways, thus benefiting as early as possible from each available source of information.

The first approach is guided by a common misconception, namely, the assumption that sensory cortical areas process information solely from the sensors they are directly connected to. In this scheme, the auditory cortex would only process auditory input sounds from the ears, while the visual cortical areas would only process visual inputs from the eyes. However, there is ample evidence nowadays that a strict separation of different modalities and the accompanying neural areas does not exist. For instance, various studies have shown the ability of the visual cortex also to process sounds (Shams et al. 2005; Iurilli et al. 2012; Vetter et al. 2014). Others have found in return that the auditory cortex can also process visual input (Sharma et al. 2000; Belin et al. 2000; Finney et al. 2001). Of course, the auditory cortex has the major role in sound processing, and the visual cortex is far from contributing as much as the former one

in sound processing. However, in the context discussed here, the question is not how important the cross-modal contribution is, but rather the fact that it *does* exist at all.

The SC is a suitable candidate for the location where cross-modal integration actually happens in the central nervous system. Perhaps nowhere is the convergence of modalities more evident than there, as asserted by Meredith and Stein (1986) based on an extensive review of research works on multimodal integration in mammal brains. Located in the brainstem, the SC is organized in seven layers, split into two functional units. One of these receives sensory inputs (mainly from vision, audition, and proprioception), the other one generates motor commands based on this sensory input. These motor commands can, for instance, be eye saccades (Moschovakis 1996), or body movements (Stein et al. 2004)—in particular head movements (May 2006).

By binding quick motor reactions to sensory inputs, the SC is thought to play an important role in attentional reactions, in particular, *exogeneous* ones.¹ Two major phenomena have been observed in attentional reactions in which the SC is involved:

- If two cross-modal stimuli are sufficiently overlapping in space and time, a synergistic effect will be observed in the multimodal neurons of the SC—a phenomenon called *multimodal enhancement*;
- This effect will be more pronounced for the stimulus of the weaker modality—a phenomenon called *inverse effectiveness* (Anastasio et al. 2000).

Moreover, multimodal integration is dependent on the *congruence* of the perceived stimuli. When two or more stimuli arise from the same perceptual entity, like an audiovisual object for instance, or when they share perceptual attributes, like an audiovisual click.² Interestingly, when there is a conflict, that is an *incongruence* between auditory and visual information supposedly belonging to the same perceptual object, vision usually takes over the other modalities, a phenomenon named *visual capture* (Hay et al. 1965). For instance, Pick et al. (1969) showed that the visual-spatial position of an object is not alterable by incongruent auditory stimuli. According to the review on *visual capture* by Posner et al. (1976), the reason why vision takes the lead on other modalities might be explained by the “relatively weak capacity of visual inputs to alert the organism to their occurrence.” Thus, attention is preferably put on visual analysis to counterbalance the relative inherent lack of saliency of visual stimuli. However, it has to be kept in mind that the relative importance of visual dominance has been reconsidered by Spence and Driver (1994, 1996, 1997a, 1997b), and later by Turatto et al. (2002). These findings are crucial for the understanding of how multimodal information is gathered and integrated. In fact, visual and auditory information are not considered equal in multimodal object formation and, consequently, concerning potential reactions to their appearance in an environment.

¹That is, reactions caused by the stimuli themselves, in opposition to *endogeneous* ones as are caused in a goal-driven way.

²An *audiovisual click* is a quick and simple sound, such as a pure tone section, presented together with a visual object, such as a dot or a cross of equal duration.

As compared to visual scenes, auditory scenes are inherently more prone to salient objects. Nevertheless, some particular cases of auditory capture over vision have been observed and reported—see Gebhard and Mowbray (1959), for instance. A later hypothesis by Welch and Warren (1980) provides a plausible explanation of the underlying mechanisms of visual or auditory capture. Obviously, vision is particularly adapted to *spatial analysis* whereas auditory fits particularly *temporal analysis*. This hypothesis, called *modality appropriateness*, is based on the specifics of the sensors.

More recently, Fendrich and Corballis (2001) used an experimental paradigm after Welch and Warren (1980) that led to the observation of a more pronounced effect of auditory capture versus visual capture. The authors of this paper have introduced the notion of *Intersensory Temporal Locking* (ITL), thus providing a more comprehensive explanation of the different observed phenomena of modal capture. The ITL, supported by a prior study of Scheier et al. (1999), is defined as a mechanism allowing the sensory cortical areas to solve potential temporal ambiguities in the perception of multimodal stimuli and offers a reasonable basis for the understanding of when either auditory or vision lead perception, and what kind of stimuli triggers such modal capture.

In addition, both the experiments in Shams et al. (2001, 2002) leading to the observation of auditory capture over visual capture, combined with the different results obtained a decade before by Saldana and Rosenblum (1993), conducted Shams et al. (2002, p. 151) to state:

The discontinuous stimulus in one modality alters the percept of the continuous stimulus in the other modality and not as strongly vice versa.

In summing up, all the studies mentioned above lead to the conclusion that multimodal perception consists of more than the sole concatenation of auditory and visual data for forming the representation of multimodal objects in higher cerebral areas. The phenomena of auditory capture, visual capture, modality appropriateness, or discontinuity versus continuity of perceived signals, indicate that audition and vision are working together closely, whereby both modalities of the two mutually benefit from this advantage.

2.2 *The Reverse-Hierarchy Theory*

Consider the cases of the voice of somebody talking in a completely silent room in contrast to talking in a very crowded and noisy place such as in a cocktail party situation. The comparison raises the following question: “*Are identical stimuli in different surroundings processed in the same way?*”

A recent model of perceptual information analysis, the *Reverse-Hierarchy Theory* (RHT), puts the following insight to the fore. The informational context in which stimuli are perceived has an impact on the deepness and thoroughness of their analysis. RHT has been introduced and put into a formalized algorithm by Hochstein

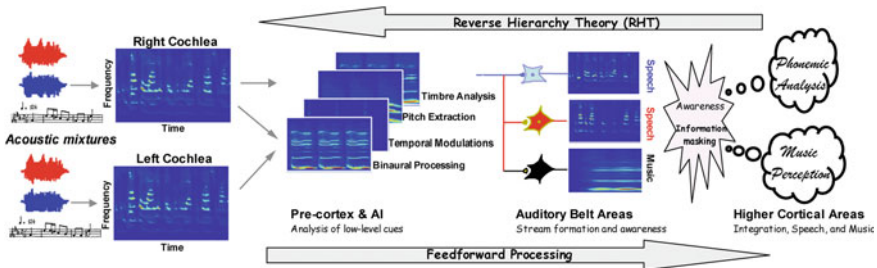


Fig. 2 Schematic of the Reverse Hierarchy Theory—after Shamma (2008)

and Ahissar (2002), Ahissar and Hochstein (2004), Nelken and Ahissar (2006), and recently Nahum et al. (2008). The core of this theory is to make the bridge between high-level representations of perceived signals, such as auditory objects, and correlated low-level cues, such as frequency spectrum, ITD, or ILD. As to the latter ones, it is of interest whether these cues are necessary or not for making high-level decisions, such as to initiate adequate motor actions. On the one hand, as low-level attributes gain in relevance for refining auditory stream analysis, the more difficult a discrimination task is, for example, for solving ambiguities. On the other hand, if the informational context is simple, the high-level representation of the perceptual streams—i.e., objects—will be directly usable, thus making deeper and more thorough analyses of the streams dispensable (Fig. 2).

The RHT is thus also linked to internal representations of the world, specifically to how perceptual streams are combined to achieve a unified and robust perception of multimodal entities, and perceptual objects. Indeed, as Shamma (2008) sums up:

If the high “objects” and their “low-levels cues” are congruent, the feed-forward process is rapid, and the use of all available salient cues is effective and comprehensive.

Thus, in addition to the capabilities of perceptual streams analysis due to powerful features extraction, the ability to rapidly provide access to high-level representations of the perceptual world is quite astonishing as well. This is due to the fact that high-level representations include temporal integration and prior assumptions about incoming sensory information—see Sect. 2.3.

Further, the RHT helps to understand attentional processes. In cases of incongruent perceptual streams such as two males speaking from close spatial positions, the theory postulates that competing cues will easily disrupt the mechanisms that require low-level cues to disambiguate the two streams. The way the sensory areas of the brain process perceptual information streams, in particular, those dealing with vision and audition, has for long been interpreted as almost exclusively being dominated by bottom-up processing. With the RHT however, there is now an innovative attempt for explaining the links between the traditional sensor-to-cortex pathway and the cortex-to-sensor one, showing that they are activated depending on the complexity of the information to be processed. Consequently, RHT is of help when constructing artificial agents equipped with human-like perception. It suggests processing the data

that such agents acquire concerning the context in which they have been collected. In particular, it shows that making assumptions about *what is coming up next* in a scene can be useful for accelerating and simplifying the processing of sensory information. To be sure, the existence of such processes in humans implies that their brain has prediction abilities. To illustrate these, the next section introduces the *Mismatch Negativity* phenomenon, a physiological reaction to deviant incoming sensory information with respect to the prediction made by these cortical areas.

2.3 *Mismatch Negativity*

This section addresses the following question: “*Can the apparition of stimuli be anticipated?*” Anticipation, or prediction, is the ability to have a strong belief about what is coming up next. This ability has the potency of considerably accelerating processing of perceptual data. Further, it enables the sensory cortical areas to detect inconsistent, salient and/or incongruent, objects. “Inconsistent, salient or incongruent” objects are such that somehow do not fit prior predictions. Consequently, they may require special reactions, such as motor commands to redirect the sensors in order to get additional data that would help understand the origin of the observed unpredictability. As an example, imagine a strong male person with an angry face uttering with high pitch and very calm voice: “*Yesterphinge*, I was in the elephant”. The following list highlights three cases in which anticipation is initiated. Yet, it may turn out to be wrong in the end.

1. The characteristics of the voice (an angry face would anticipate a loud, low-pitched voice)
2. The semantic content of the speech, that is, certain words have a higher probability of occurring in the given context (“...in the elephant”)
3. The words themselves, given the context and the initial syllables (“*Yester-phinge*” instead of “*Yesterday*”).

For all three cases the following holds. If what is perceived does not match prior expectation, a quick reaction is triggered. One of the first reactions to these unexpected perceptual objects occurring in a predictable stream of information can be observed in the sensory areas, such as the auditory or visual cortical areas, in terms of a particular neuronal response, the *Mismatch Negativity* (MMN).

This effect, when elicited, signals a quick attentional response to objects that do not match the expectations of the sensory areas. Discovered by Näätänen et al. (1978), the MMN can thus be described as a quick, specific reaction to the *incongruence* of an auditory or visual object concerning the short-term context in which it appears. MMN is mainly present in the auditory areas (Molholm et al. 2005), specifically in the temporal superior cortex and the frontal cortex (Alho 1995). It occurs at around 100–200 ms after the deviant stimulus. For instance, when in a repeated sequence of sounds of a center-frequency of 1000 Hz, unexpectedly a sound at 1032 Hz is presented, it will be recognized as deviant from the predictable sequence perceived

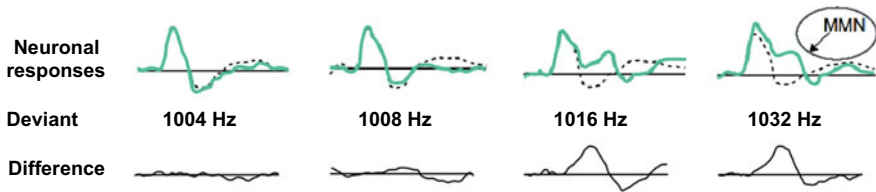


Fig. 3 *Mismatch negativity*. **Upper curve:** Neural responses recorded in 80% of the occurrences of randomly presented sounds of 1000 Hz center frequency (**black dotted lines**), of deviant sounds at different center frequencies in 20% of the occurrences (**green lines**). **Lower curve:** Differences of the responses to deviant sounds as compared to the 1000 Hz reference—after Näätänen et al. (2007)

so far. The neuronal reaction to this deviant sound will show up as the MMN—see Fig. 3. MMN has also been observed when there are amplitude or timbre variations (Näätänen and Alho 1995). It is thus an effect linked either to the apparition of new percepts or to variations in the perceptual attributes of ongoing ones.

Mismatch negativity is undoubtedly an indication of a reaction to an unpredictable stimulus. Yet, its role in the formation of a perceptual world model has also to be seen under the following aspects. By being able based on only three or four occurrences of a stimulus, to infer a rule that enables the prediction of the next stimulus to appear, the sensory cortical areas can speed up the processing of the incoming stream of stimuli by just checking if the perceived stimulus matches the prediction. If it does match, there is no need to process the stimulus fully, and computation time is saved (behavior to be linked to the RHT, see above). However, if it does, a warning signal, the MMN, is generated to potentially initiate a motor reaction such as a head movement. This reaction is a way to motivate a more in-depth analysis of the unpredictable stimulus, for instance, by bringing other available and relevant sensors into play that gather additional information for the analysis.

Friston (2005) has highlighted the fact that the brain’s internal representations of the world can be utilized to predict what most probably happens next in the environment. Along this line of thinking, Lochmann and Deneve (2011) introduced the notion of *predictive coding* for causing inference of sensory objects that are not directly recognizable from sensory cues. Arnal and Giraud (2012), in their review of cortical oscillations and sensory predictions, listed several mechanisms that allow the auditory cortex to predict the point in time when a stimulus is most likely to happen in the given context. In fact, MMN accompanies all prediction processes in the brain. Yet, for two reasons these processes are more than just pure anticipation: (i) it analyzes the perceptual scene faster and, (ii) it represents a powerful way of revealing unpredictable changes in the perceptual stream of information, especially in the auditory one.

Mismatch Negativity teaches us that when auditory (and visual) stimuli are processed, the sensory cortical areas are able to form a *predictable sequence* quickly, thus enabling instant detection of perceptual irregularities. Already at this stage of sensory information processing, the MMN reveals that there is no stimulus standing

out per se: a stimulus can be detected as deviant, or *incongruent*, in a certain sequence of perceptual objects, but would be rated as “normal” in another sequence where this stimulus would be surrounded by similar ones. Consequently, the likelihood of a stimulus to trigger an attentional reaction has to be defined in relation to its environmental surroundings. For sure, these surroundings can be variable. Thus, stimuli may change their perceptual role accordingly. They will elicit a different behavioral response from one situation to another one, from one context to another, from one place to another, and/or from one point in time to another one. Whereas the MMN has not yet been linked directly to any direct motor reaction, its strong involvement in attentional reactions (Escera et al. 1998, 2003) makes it a solid candidate for triggering eye, body, or head movement in the presence of incongruent stimuli or perceptual objects, especially through the notion of *saliency*.

2.4 Saliency

Saliency is a measure of how much a stimulus, such as a sound wave or the pixel of an image, differs from its surroundings, be it temporally or spatially. In human perception, saliency has mainly been studied in vision. In particular, following the definition of Treisman and Gelade (1980), saliency stems from local singularities that are exhibited within a stream of perceived data. For instance, within an image composed of numerous red circles, the presence of a unique green one would present a local singularity in terms of color: the green circle would then be considered as salient. From this analysis of the perceptual streams, and mainly exhibited by the auditory and visual cortical areas, attentional reactions can be elicited, such as eyes movements towards visual stimuli of high intensity (Wolfe 1994; Nothdurft 2006). Moreover, saliency is shaped and influenced by learning and experience. For instance, while a musician can detect a false note instantly without even having to focus on listening, it could remain unnoticed by an untrained person. In the visual system, the primary visual cortex (V1) already has a map of visual saliency (Li 2002). Mazer and Gallant (2003) have shown that the activity of neurons of the extrastriate visual area (V4), a structure placed higher in the hierarchy of visual signals analysis, can predict towards which particular area in space an eye saccade will be directed on an ongoing visual exploration task. This observation supports the assumption of the presence of a topographical map of saliency in V4. Further, the intra-parietal lateral area (Bisley and Goldberg 2006) and the frontal eye field (Thompson and Bichot 2005) have been associated with the phenomenon of visual saliency as well. The human auditory system also responds well to saliency, and potentially also triggering motor reactions, in particular, head and body movements. However, the attributes that the auditory sense is sensitive to, and on which it bases its interpretation of the auditory scene, are different from those used in vision.

As concerns saliency, the auditory system mainly processes spectral and temporal modulations (Yost 1992; Alain et al. 2001) and, based on these, it can extract auditory entities of relevance even in noisy environments (Hall et al. 1984). Addressed acous-

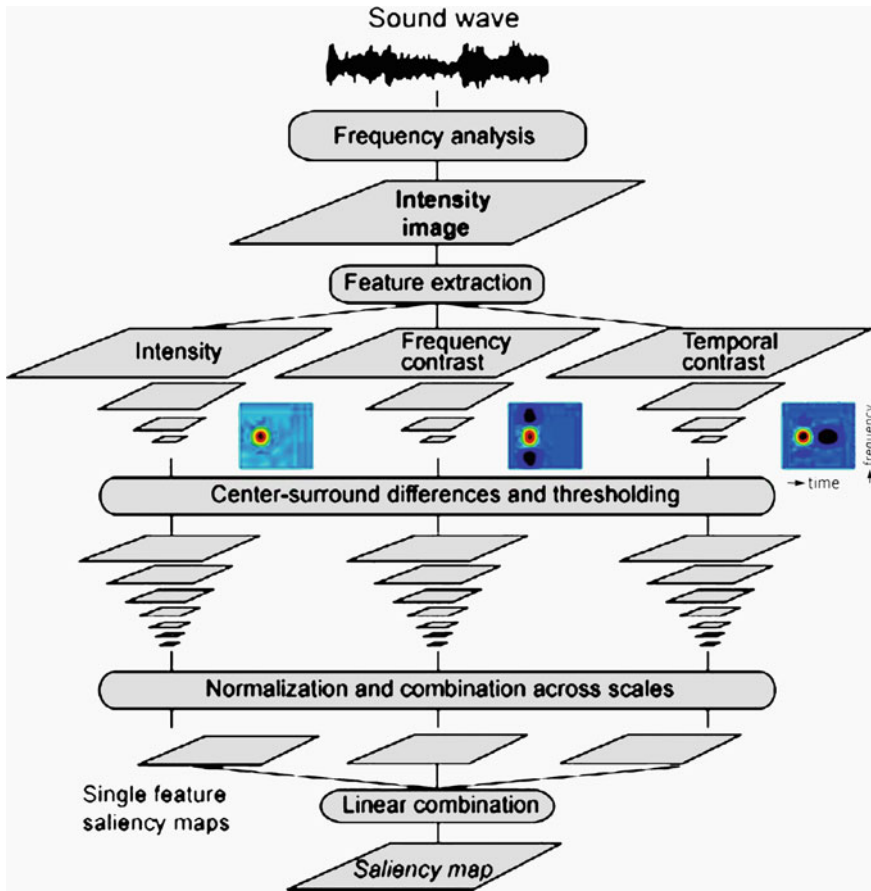


Fig. 4 Building an auditory-saliency map after Kayser et al. (2005), inspired by the work of Koch and Ullman (1985). An acoustic wave is received and then converted into a spectro-temporal representation allowing the extraction of attributes such as intensity, spectral and temporal contrasts. The resulting maps are combined into a comprehensive auditory-saliency map following a normalization step

tic attributes are predominately spectral contrast, temporal contrast, and intensity. These are then exploited in parallel by neurons of the auditory areas, consequently leading the formation of saliency maps dedicated to specific attributes. Then, these maps are merged in order to create a global map of auditory saliency of the actual acoustical environment—compare Fig. 4. Be it for the visual or the auditory system, the creation of saliency maps within dedicated sensory areas is an important step toward understanding the internal world representations of each of these systems. Indeed, these maps provide potential candidates for a reorientation of attention. For instance, a person speaking from a specific position outside the visual field of a listener requires head movement, or a suddenly moving target demands an eye saccade.

Saliency, be it visual, auditory, or multimodal, has intensively been discussed, modeled and implemented in the robotics, artificial intelligence, and computational neurosciences communities—see for instance Koch and Ullman (1985), Itti et al. (1998), Oliva et al. (2003), Kayser et al. (2005), Duangudom and Anderson (2007), Ruesch et al. (2008). While saliency is a key component to demonstrate how low-level attributes shape motor reactions, current concepts are not sufficient for a comprehensive understanding of attention reorientation. Indeed, by solely considering low-level attributes of the perceived stimuli, the means for including feedback from higher levels of the central nervous system are rather limited, though not impossible—compare Blauert and Brown (2020), this volume.

2.5 Conclusion of Sect. 2

This section discusses how the sensory areas of the brain deal with information coming from different sensors, each one having its own very particular characteristics, to trigger relevant behavioral reactions to the incoming perceptual streams. This question was addressed by presenting four phenomena that are a small part of the global and complex mechanism of *attention*: (i) the *Superior Colliculus* as a brain structure responsible for multimodal integration and consequent motor reactions, (ii) the *Reverse Hierarchy Theory* as an attempt to explain how the sensory areas compute stimuli differently given their level of ambiguity and the specific surroundings, (iii) *Mismatch Negativity* as a quick neuronal response to localize unpredictable perceptual objects, and (iv) *Saliency*, as a reaction to local singularities low-level characteristics of perceived signals are susceptible to exhibit. Each of these phenomena represents an important part of attention in multimodal perception. Integration, prediction abilities, detection of incongruences, selective in-depth analyses of the perceptual streams, and motor reactions are directly bound to each of these neuronal phenomena. The *active* component of perception is particularly relevant in this context. Indeed, whenever there are ambiguities in the understanding of an environment, motor reactions will enable the brain to access new information for refining its previous representation of the scene. In doing so, this additional information will help to solve the previous ambiguities. At the same time, learning mechanisms will continuously increase the system's knowledge, and thus prepare the system for similar future tasks. For instance, when the position of an auditory object seems to be *odd*, that is, incongruent or unexpected, turning the head toward this object will initiate the redirection of visual sensors to an adequate position for better localization.

The next section will introduce a computational model rooted in the biological phenomena described here. It provides attentional behavior for a mobile robot.

3 Modulating the Head Movement—The HTM Model

The previous section listed and described important mechanisms playing a role in attention, perception, and motor reactions to either incongruent, salient, or unpredictable events. From a technological point of view, several attentional systems have already been implemented. Most of them, however, share an important feature: they heavily rely on data that have been gathered before the robot even started its life, data for which dedicated learning systems have been trained to solve very specific problems that the robot has not even encountered yet. When considering the phenomena presented above, none of them rely on prior learning of specific skills. Saliency is a property of the signals, MMN is a very short-term reaction and is profoundly adaptable, so as the RHT. The SC computes a quick multimodal integration directly followed by a motor reaction depending on the content of the incoming multimodal information. Consequently, it should be possible to design an artificial system that implements the key features of human auditory (and visual) attention without having to gather a vast amount of training data in advance that could help solve only one or few specific problems.

This section will thus describe the Head-Turning-Modulation model, a model aiming at providing an answer to the central question of this chapter, which asks: *How can audition be used as a trigger for head movements towards objects of interest?* In this context, three important aspects were presented with regard to the global phenomenon, *attention reorientation*, in which the question mentioned above is included. In the following, an attempt is described to provide a binaural and binocular humanoid robot with the ability to learn how to identify unexpected auditory objects and, when appropriate, trigger head movements toward these objects for collecting supporting visual information. This model of high-level attention, recently introduced by the authors in Cohen-L'hyver et al. (2015, 2016), Cohen-L'hyver (2017), Cohen-L'hyver et al. (2018) is mainly based on the four biological phenomena already discussed. The main contributions of the HTM model are outlined here with a specific idea in mind, i.e. the characteristic behavior of artificial agents can be achieved without having to deal with overly complex algorithms.

The section is organized as follows. The first part is dedicated to the description of the concepts that the HTM relies on, that is, especially the two modules that constitute it: (i) the Dynamic Weighting model, and (ii) the Multimodal Fusion & Interference module. The second part introduces aspects of algorithmic formalization of the two different modules. Finally, a third part presents some of the results obtained in simulations and on a real robot.

3.1 Concepts and Global Architecture

The Head-Turning Modulation acts similarly to a Blackboard system (Schymura and Kolossa 2020, this volume) and contains two primary modules³:

- The *Dynamic Weighting* module (DW) is deciding whether an audiovisual object appearing in the environment is *incongruent*, given the other audiovisual objects already detected in the past in this environment,
- The *Multimodal Fusion & Inference* module (MFI) is in charge of providing the DW module with corrected and completed *audiovisual classes* as a basis for the computation of congruence.

As shown on Fig. 5, the HTM exploits auditory and visual labels provided by dedicated classification experts for emitting hypotheses (i) on the *audiovisual class* the detected sources belong to and, (ii) on which of these sources the robot should focus. Each computational expert is dedicated to the detection and the recognition of particular *auditory labels* or *visual labels* (Two!Ears et al. 2012). For instance, one expert deals with the detection and recognition of the sound `speech`. Another one is assigned to the sound `barking`, and still another one addresses the visual entity `male`, and so on. On this basis, each hypothesis might potentially lead to the triggering of head movements towards audiovisual sources of interest.

Importantly, these audiovisual sources appear randomly in the environment—that is, by not following any pattern the robot either understands or not and, consequently, can predict or not. By the way, triggering head movements towards any audiovisual source would not require any form of particular intelligence. The low-level attributes of the signals are often sufficient to localize the objects for sending meaningful motor commands. Here, however, the goal is to *modulate head movements*, that is, to either trigger *and* inhibit them. Indeed, not all of these head movements are relevant. For instance, turning the head toward the tenth `barking dog` in a room populated with only barking dogs, is very likely redundant such as not providing any useful additional information. Thus, by inhibiting some head movements, the head of the robot can be used for other kinds of movements, as may be requested by other tasks. The two modules constituting the HTM module have been designed and implemented in a way that they can understand the environment being explored by the robot in terms of audiovisual objects of *importance*. Thereby, the attribute of importance is assigned to objects in the following ways.

- The DW module implements the notion of importance through the concept of *congruence*. Congruence is defined here as *semantic saliency* since it is not applied to the low-level attributes of the perceived signals, such as spectral composition, ILD, or ITD, but rather on high-level representations of these signals, namely audiovisual classes. The audiovisual classes $c(a, v)$ are made by the concatenation of an auditory label, a , and a visual label, v . On this basis, and without any prior knowledge of the actual environment, the DW aims at determining whether an

³These are called “Knowledge Sources” in the integrated TWO!EARS software.

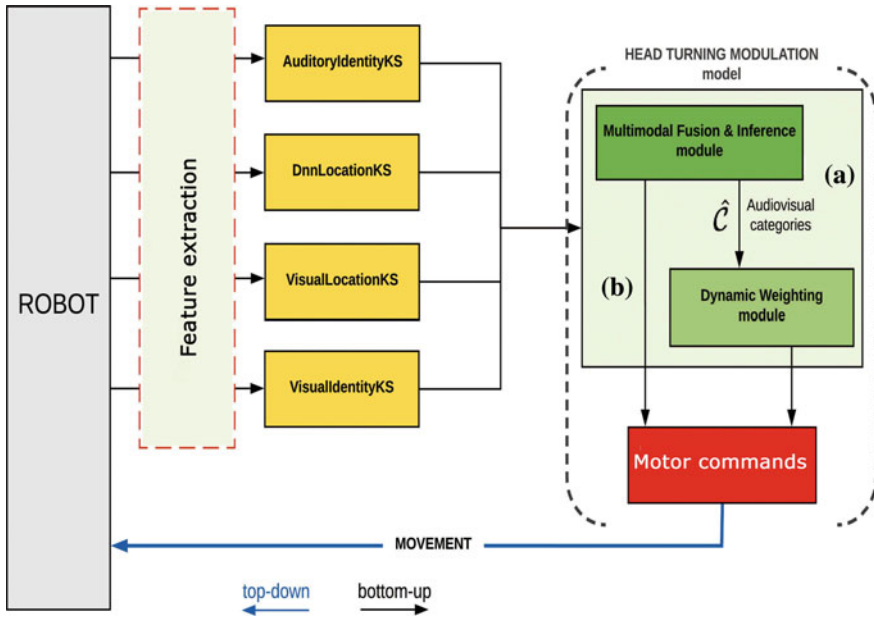


Fig. 5 Schematic architecture of the HTM model and its two main components: **a** The Multimodal Fusion & Inference module is in charge of providing to the DW module corrected audiovisual classes from classification experts outputs. **b** The Dynamic Weighting module estimates the *congruency* of an audiovisual object, in its current environment. Each of the two modules can trigger head movements separately. The **red box** depicts the computational component that integrates different motor commands and puts them into an order to prioritize one of them, depending on the actual situation

audiovisual source is *incongruent* or not in the environment being explored. If it is, a motor reaction is triggered toward this audiovisual object. This motor reaction can be compared to those triggered by the Superior Colliculus (see Sect. 2.1) or the MMN (see Sect. 2.3).

- The MFI module implements the notion of importance through the prism of *reduction of uncertainty* of the auditory and visual labels received from the classification experts. More precisely, the MFI module analyzes the uncertainty that it senses with regard to the combination of auditory and visual information, in particular, regarding the assignment of *audiovisual labels*, a combination that contributes to the multimodal representation of objects as used within the HTM. The MFI module is primarily inspired by the Reverse Hierarchy Theory (see Sect. 2.2). A further aspect originates from the principle of *intrinsic motivation* of a person or an artificial agent to accomplish a particular action for the sake of an internal rewarding system, such as the one Berlyne (1950, 1954) has first described and theorized. Compare also Macedo and Cardoso (2001), Baranes and Oudeyer (2009, 2010) for examples of artificial systems furnished with such kind of motivations. In practice, the classification experts mentioned above are not unlikely to provide erroneous

labels due to classification errors or even missing labels due to occlusions, such as happens when objects are placed outside the field of view of the robot. Thus, the MFI module, being directly coupled with DW module, is in charge of providing the estimated audiovisual classes, $\hat{c}(o_j)$, the perceived objects o_j might belong to. This analysis consists of a fusion of auditory and visual information as acquired by an active unsupervised learning algorithm, linked to the usage of head movements.

One potential issue arises here; that is, both modules have the ability to trigger head movements for their respective task. The MFI generates motor commands to acquire its multimodal representation, while the DW generates its commands for an attention-driven behavior directed to incongruent objects. Both head movements must then be assigned priorities—see the red box in Fig. 5. Since the DW makes decision based on congruence of perceived audiovisual objects, this information must be exempted from any classification or fusion errors. Thus, the motor commands triggered by MFI are prioritized against the ones triggered by DW. The following subsection provides details about the two modules constituting the HTM.

3.2 Algorithmic Formalization

This section provides details of the algorithmic formalization of the two modules constituting the HTM, modules that respectively rely on *congruence* and intrinsic motivation through reduction of uncertainty. As mentioned before, the HTM relies on the notion of *multimodal object* populating the environments the robot will explore. However, this notion of an object is not objective: it is already an interpreted notion arising from the convergence of different streams of information into a unified and coherent internal representation. Thus, considering that the environments are objectively populated with audiovisual *sources* emitting auditory, visual or audiovisual *events* Ψ_k , one of the first task of the HTM is to emerge the notion of object, such as

$$\Psi_k = \{\theta_k, c(\Psi_k)\} \longrightarrow o_j = \{\hat{\theta}_j, \hat{c}(o_j)\}, \text{ with } \hat{c}(o_j) = \{\hat{c}^a(o_j), \hat{c}^v(o_j)\}, \quad (1)$$

where c represents the real audiovisual class of the event Ψ_k , \hat{c} depicts the estimated classes (audio, visual or both) the object o_j belongs to, θ_k the real angular position of the event, and $\hat{\theta}_j$ the estimated one by the localization expert. The estimated classes \hat{c} come from the analysis performed by the MFI (see Sect. 3.2.2) of the data brought by the audio and visual classification experts that have been trained beforehand to identify particular sounds and images. Also, it is these audiovisual classes that will be utilized by the DW to compute the congruence of the concerned object. The raw data the HTM will retrieve at time step t from the Blackboard system will be organized as follows:

$$\mathbf{V}[t] = (\mathbf{P}[t], \mathbf{\Theta}[t]), \text{ with } \mathbf{P}[t] = (\mathbf{P}^a[t], \mathbf{P}^v[t]) \text{ and } \mathbf{\Theta}[t] = (\mathbf{\Theta}^a[t], \mathbf{\Theta}^v[t]), \quad (2)$$

where on the one hand $\mathbf{P}^a[t] = (p_1^a[t], \dots, p_{N_a}^a[t])$ and $\mathbf{P}^v[t] = (p_1^v[t], \dots, p_{N_v}^v[t])$ are the vectors of probabilities from the auditory and the visual classification experts, and $\Theta^a[t] = (\theta_1^a[t], \dots, \theta_{N_a}^a[t])$ and $\Theta^v[t] = (\theta_1^v[t], \dots, \theta_{N_v}^v[t])$ are the vectors of probabilities from the auditory and visual localization experts, respectively. These are precisely the vectors \mathbf{P}^a and \mathbf{P}^v that the MFI will try to correct or, whenever one of them is missing, to infer.

The following section introduces the DW corresponding to the highest level, that is the closest to the cognitive abilities of the HTM.

3.2.1 Congruence—The DW Module

Within the DW, the emphasis has been put on dealing only with high-level representations of the perceived multimodal data, namely the auditory classes they belong to. Following this idea, the aim of a reactive robot—in terms of head movements as driven by the concept of congruency—is the detection of incongruent audiovisual objects in an unknown environment in comparison to prior observations. The system has neither access to the content of the multimodal objects that populate this environment nor to their time of appearance. The only tool that the HTM has when entering a new room is a set of classification experts that have been trained beforehand.⁴ Further, the DW is designed to exploit additionally available knowledge for future use in unknown environments.

Congruence is based on conditional pseudo-probabilities where the probability of observing a certain audiovisual class depends on the environment in which it occurs. In other words, the less often an audiovisual object has been observed in the past, the less likely it is to occur again in the future.⁵ On the contrary, the more frequently an audiovisual object has been observed in the past, the more likely it is to occur again in the future.

This has been formalized by means of the posterior probability of an object o_i to belong to a class $c^{(l)}(a_i, v_k)$ in the l th environment $e^{(l)}$:

$$p(o_j \in c^{(l)}(a_i, v_k) | e^{(l)}) = p(c^{(l)}(a_i, v_k) | e^{(l)}) = \frac{|c^{(l)}(a_i, v_k)|}{N_l}, \quad (3)$$

where $|c^{(l)}(a_i, v_k)|$ depicts the number of objects that have already been associated to the audiovisual class $c^{(l)}(a_i, v_k)$, and N_l is the total number of objects detected so far. Since no information is available about what class is more likely to occur in a given environment, the probability $p(o_j \in c^{(l)}(a_i, v_k) | e^{(l)})$ will be compared to the equiprobability $K_l = 1/|C^{(l)}|$ of observing any class detected so far. Thus, it is possible to take a decision on the congruence of the considered object by

⁴These have been provided by the TWO!EARS project software freely available from www.twoears.eu.

⁵Obviously, this indicates a link to Bayesian theory.

$$o_j \in c^{(l)}(a_i, v_k) \text{ is incongruent} \Leftrightarrow p(c^{(l)}(a_i, v_k)) \leq K_l. \quad (4)$$

Following that, and to render the notion of *importance* of the emitting objects, two functions have been designed to assign their weights w_{o_j} with respect to their congruence

$$w_{o_j}[n] = \begin{cases} f_{\omega}^{\bullet}[n] = 1/(1 + 100 e^{-2n}) & \text{if } p(c^{(l)}(a_i, v_k)) \leq K_l \\ f_{\omega}^{\circ}[n] = (1/1 + 0.01 e^{2n}) - 1 & \text{else} \end{cases}, \quad (5)$$

where f_{ω}^{\bullet} is an increasing positive function converging to 1 and dedicated to incongruent objects (high weight equals high importance), f_{ω}° is its symmetrically decreasing negative function converging to -1 and dedicated to congruent objects, and where n is a temporal index that is systematically reset to zero whenever the congruence state of the object changes. To trigger a head movement, the object with the highest weight, that is, the most incongruent, will be considered as the target of the motor reaction. Also, if two objects share the same weight, the one that appeared the latest would be prioritized, thus applying a form of motivation by *novelty*. Note that the computation of the motor orders, not detailed here—see Cohen-L’hyver (2017), and Cohen-L’hyver et al. (2018) for complete description, is conceptually and mathematically formalized by the use of a GPR model—developed by Gurney et al. (2001a, b) and inspired by the basal ganglia-thalamus-cortex loop present in humans and playing an essential role in motor command selection.

All of this leads to the very definition of *environment*. The robotics community defines an environment most often by its physical existence, or its topographical characteristics: size of the rooms, number of access points, usable paths, zones of danger, lighting conditions, e.g., see Makarenko et al. (2002), Durrant-Whyte and Bailey (2006), Cuperlier et al. (2007), and Baranes and Oudeyer (2010). In the context of the DW, and thus of the whole HTM, an environment is also defined through a *semantic* approach, namely, by the audiovisual objects, or entities, that are present in it. Going even a bit further, a refined definition reads as follows:

An environment is defined by the relative congruence of all the audiovisual classes that have been perceived in it.

In the vein of this definition, two very different rooms, such as two conference rooms at different universities, will be considered as being identical, if and only they share the same set of audiovisual classes congruence values. Consequently, the respective status of congruence of the audiovisual classes detected in all the already explored environments consequently constitute the *knowledge* of the world the DW creates. This knowledge is used by the DW whenever it detects that the current explored environment is similar enough to one the robot already explored in the past. Being able to transfer acquired knowledge to new unknown environments quickens the understanding of it by taking advantage of the experience of the robot.

However, taken that congruence relies on a multimodal representation of the objects perceived in the explored environments, what happens when an object is

placed behind the robot, thus hindering it from acquiring adequate visual information? Turning the head toward the object to get the full data in order to accurately compute congruence would definitely be absurd, if this object would be thereafter considered as congruent, a head turn would have already been triggered. Such conflicting situation motivated the creation of a Multimodal Fusion & Inference module, as described in the following section.

3.2.2 Reduction of Uncertainty—The MFI Module

A second module with the ability of *inferring missing data* has been developed to circumvent a deadlock situation of the mentioned kind. The Multimodal Fusion & Inference (MFI) module also constitutes a reflective feedback loop that uses auditory and visual data coming from the sensors (after the dedicated classification experts have processed them) to send back a motor command, as illustrated in Fig. 6. This motor command will give the robot access to new data that might redefine the best motor action for the robot. Since the system relies solely upon a high-level representation of the perceived data, namely, audiovisual classes, the inference made by the MFI module will be about auditory labels given known visual ones, or about visual labels given known auditory ones.

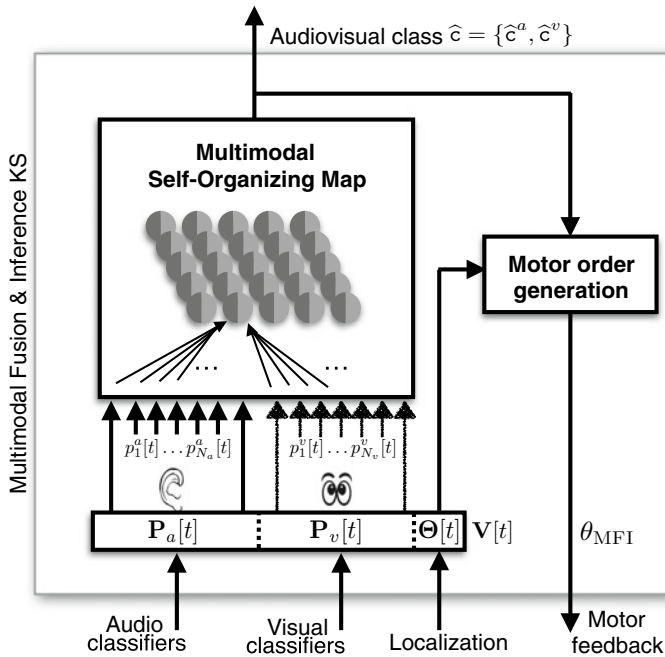


Fig. 6 Multimodal Fusion & Inference module architecture

It is necessary to learn the relationships between auditory and visual labels to achieve this—such as `barking dog` or a `speaking male`. In other words, every time the robot faces an object which emits sound, the MFI module will take the chance to learn the audiovisual pair that is perceived. Once this learning has been accomplished, the MFI can offer an inference of a missing modality. However, it has to be kept in mind that classification experts are prone to errors—particularly the auditory experts when the acoustic conditions become challenging, such as in reverberant and noisy surroundings, or when the explored environment differs too much from the one used for the prior experts training. Thus, relying too much on the output of these classifiers would lead to erroneous learning of the audiovisual pairs.

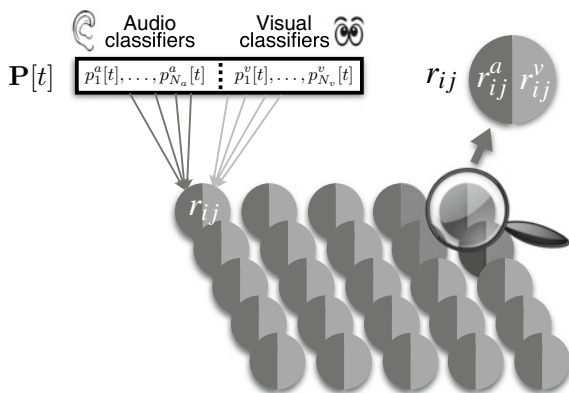
The MFI has been designed around a *Self-Organizing Map* (SOM), after Kohonen (1982). Such learning algorithm performs a vector quantization of high-dimensional input data into a lower dimensional map (in our case, two-dimensional). Indeed, a SOM is a map composed of a certain number of nodes (or neurons) that represent the constituting vectors of the matrix of data to be processed. A SOM organizes these vectors in space by assigning them a particular node within the map. What results from this procedure is a modified representation of the input data as a map that has a lower dimension than the initial set of vectors, making it easier to process while also enabling the categorization of the input data. The SOM map is *tonotopically* organized. This means that when two regions of the SOM map are spatially close, the data that they represent are also close. The purposes of an SOM are organizing the existing data in clusters, then determining the class that a new input belongs to by localizing the node within the map which is most similar to the new vector, and finally, identifying the cluster that this node belongs to. However, while the SOM algorithm provides a powerful unsupervised learning paradigm, it had to be adapted to the particular conditions in which the HTM, and the MFI in particular, have access to the data it has to process.⁶

The first major change comes from the use of not only one SOM to learn the data, but of one SOM *per modality* used to define an object, thus creating the *Multimodal Self-Organizing Map* (M-SOM), as depicted in Fig. 7. Here, auditory and visual data have been used to define an object. The overall M-SOM used in the MFI thus includes two interconnected subnetworks that will jointly participate in the creation of the internal representation of the robot's world, in terms of the audiovisual classes that have been observed during its exploration, as Fig. 8 illustrates.

The second major change consists of modifying the learning process. Indeed, while the SOM is built, and usually used, to process full matrix of data, and since, as already stated before, the HTM does not have access to prior knowledge about the objects appearing in the environments, the M-SOM will only be fed with one vector of data at a time. That is, whenever a vector of data is available, the MFI has to be capable to integrate it in the M-SOM so that a learning iteration can happen. Since the goal of the MFI is to learn the relationship between the two modalities, a vector

⁶It will only be presented in this chapter what has been changed conceptually. See Cohen-L'hyver (2017), Cohen-L'hyver et al. (2018) for a thorough description of all the contributions of the M-SOM.

Fig. 7 Illustration of the Multimodal Self-Organizing Map (M-SOM) which embeds two subnetworks, each of them being dedicated to coding the information from each modality used to define an object—audition and vision in this case



of data is sent to the M-SOM if and only if these data come from both visual and auditory sensors and are about the same object, that is, whenever the robot faces an object emitting sound. Moreover, it is important here that the reflective feedback loop is present, through the triggering of head movements towards audiovisual sources of interest, in order for the robot to face these sources belonging to audiovisual classes that might need further learning.

Third, while in a traditional SOM the proofs of convergence are numerous, the problem the MFI has to solve does not simply imply one, or several, good solutions. Since the robot is being always designed to explore unknown environments, there is no possibility to know what are all the audiovisual classes that will be present. Consequently, the MFI implements the notion of *local convergence* of the M-SOM. In particular, the quality of learning will be assessed by the MFI on a *class-by-class* manner: if the estimation of the audiovisual class an object is supposed to belong to is not trustworthy enough, more audiovisual data will be required in order to enhance the quality of the knowledge about this class. Such additional data is obtained by triggering a head movement towards the concerned source. Local convergence is formalized by the implementation of an inference ratio of $q(c^{(l)}(a_i, v_k))$ to determine whether an audiovisual class, $c^{(l)}(a_i, v_k)$, needs to be further learned by the M-SOM in an environment, $e^{(l)}$, or whether it has converged already to a trustworthy representation, according to:

$$q(c^{(l)}(a_i, v_k)) = \frac{\sum_{n=1}^{n=t} \delta_{i,k}^{\text{miss}}[n-1] \delta_{i,k}^{\text{all}}[n]}{\sum_{n=1}^{n=t} \delta_{i,k}^{\text{miss}}[n]}, \tag{6}$$

$$\text{with } \delta_{i,k}^{\text{all/miss}} = \begin{cases} 1 & \text{if } \hat{c}^{\text{all/miss}}(o_j) = \{a_i, v_k\}, \\ 0 & \text{else.} \end{cases}$$

Equation(6) describes the behavior of the MFI when it comes to setting up a hypothesis about a missing modality, and this hypothesis constitutes the reflective core of the feedback loop the MFI represents. If the ratio is too low, a command

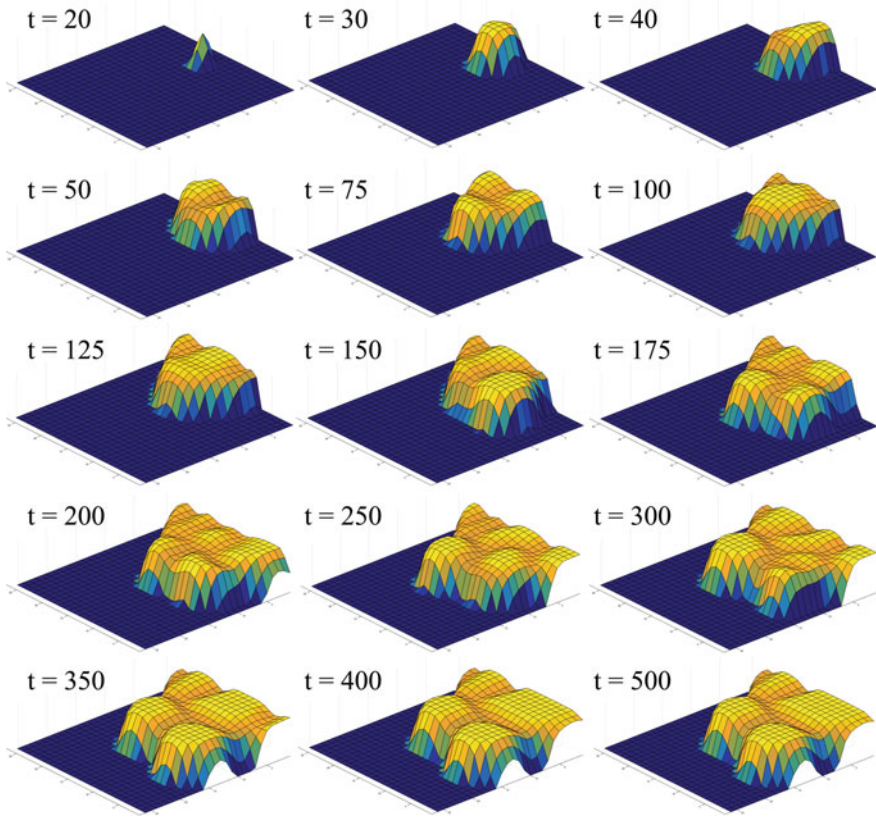


Fig. 8 *Multimodal Self-Organizing Map (M-SOM)*. Each square represents a node (or a neuron) that codes a particular distribution of the input data to be analyzed. The figure shows the evolution of such a map during 500 time steps in an experiment in simulated conditions. In the beginning, the map is unorganized, and gradually, with the amount of data it is fed with, it creates clusters of neurons that represent similar categories of data. This M-SOM embeds two interconnected SOMs dedicated to each modality used to define the notion of an object (audition and vision here). Four audiovisual classes have been created here, as the four highest regions of this map depict. The M-SOM is used after that to find the class of a new vector of classification experts data—after Cohen-L’hyver (2017), and Cohen-L’hyver et al. (2018)

will be requested for turning the head toward the sound source in order to acquire visual data. By doing so at time $t + 1$, the inference ratio $q(c^{(l)}(a_i, v_k))$ will be updated with the new information and used to feed the M-SOM thus refining the learning—given that the data from the missing modality is now available. This ratio will then be compared to a dynamically changeable threshold $K_q \in \mathcal{R}^+ = [0, 1]$ to decide whether it is now high enough to accept the inference as trustworthy. If the answer is “yes”, no head movement will be initiated. If the outcome is “no”, a head movement will be triggered. The threshold affects how quickly the MFI trusts its inference abilities.

For instance, a threshold of 0.2 would allow for eight out of ten wrong inferences on a particular class before stipulating that the inference is not trustworthy. Likewise, a threshold of 0.9 would require at least nine on ten useful inferences before inhibiting head movements. The presence of such a threshold may suggest that it is solely responsible for the global performances of the MFI, but this not the case as it is explained later in this section. Extensive evaluation of the impact of the threshold value on the quality of MFI knowledge has revealed that variations are low for threshold values in the range of 0.5–0.9 (Cohen-L'hyver 2017).

For this reason, the option of setting different threshold values matters. The lower the threshold, the fewer head movements will be triggered, but potentially more errors will be made. On the other hand, the higher the threshold, the more head movements will be triggered. Consequently, a suitable adaptation of the threshold can make sense when considering the specific situation that a robot is exposed to. For example, in a search-and-rescue scenario the priority would be put on the search for victims, thus not requiring a full understanding of all audiovisual entities that are present in the current environment (low threshold), while in a room without any high priority task to accomplish, the robot has all the time needed for a complete exploration (high threshold).

Concerning the computation of motor orders potentially triggered by the MFI, it has been formalized similarly to the DW (see Sect. 3.2.1), that is through a GPR model enabling the selection of which object needs to be focused on.

To sum up, the main purpose of the MFI is the reduction of uncertainty by using motor reactions, hence implementing a reflective feedback loop that links information from classification experts to a motor command that will in return provoke the perception of new data, and so on. Therefore, two hypotheses are set up concerning whether an audiovisual object belongs to a specific class, in particular, to one that is based on the incoming stream of auditory labels and to another one addressing the stream of visual labels. As an example, if the robot faces a person and perceives a barking sound originating from the same location: how confident would the MFI be that this audiovisual source belongs to the audiovisual class `barking person`? The possible behavior of the MFI in such a case may alternatively be as follows:

1. The robot has encountered several (`barking person`) in the past, and the MFI is now confident that it is not a classification error. The DW module can thus rely on this audiovisual fusion for computing the congruence of this audiovisual object.
2. The robot has never encountered such an audiovisual class and will thus need to gather further auditory and visual data before potentially creating a new audiovisual class.
3. The robot has already encountered this class but is still not confident enough to determine that the source does indeed belong to it. In this case, the MFI will initiate a head movement to gather more auditory and visual information.

3.2.3 Combination of the Two Modules

The combination of the DW and MFI modules consisted mainly in dealing with which module should take the lead whenever both are triggering a head movement. Still staying within the paradigm of the GPR implementation of motor commands (see Sect. 3.2.1), the computation of the motor orders triggered by the DW has been slightly modified in order to take into account the activity of the MFI, so that, in fine, the MFI is prioritized over the DW. Indeed, the former being dedicated to providing the latter with clean data, it has to take over the DW until the MFI is confident enough in its knowledge (see 3.2.2). Combining the two modules leads to a global behavior of the HTM in three phases, as depicted in Fig. 9, in a simulated environment (the time here thus corresponds to discrete time steps). At first the MFI is prioritized until $t = 135$ time steps, since it is gathering information and creating knowledge. Then, from $t = 135$ to $t = 310$ time steps, both modules trigger head movements: the MFI is confident in its knowledge about certain audiovisual classes (*speech male* for instance) but not about others (*crying female*). Finally, from $t = 310$ time steps to the end of the simulation, the MFI does not trigger any head movement letting the DW in sole charge of deciding of the importance of the audiovisual objects present in the environment.

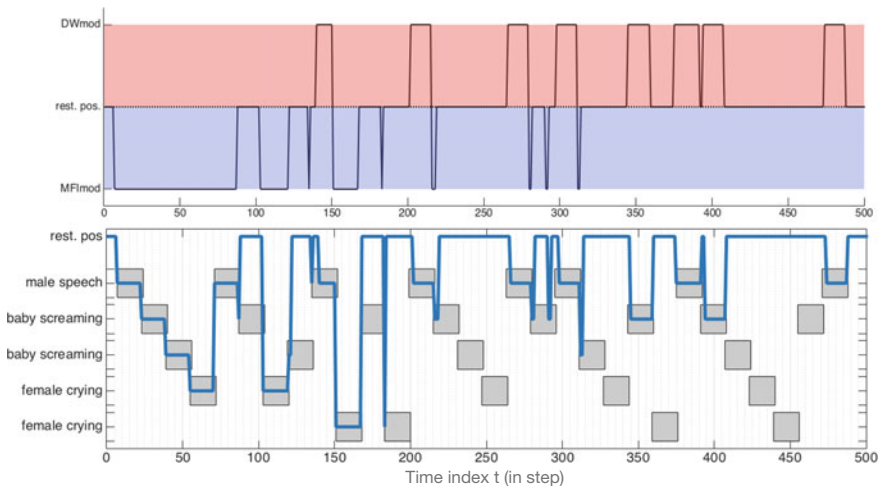


Fig. 9 Three-phase behavior resulting from the combination of the DW and the MFI in an exploration task. The x-axis depicts time steps (simulated scenario). **Top:** Head movements triggered by the DW (**up, red**) or the MFI (**down, blue**). **Bottom:** Time course of the scenario depicting which and when audiovisual objects are appearing in the environment. The black line denotes the object to which the robot drives its attention

3.3 Experiments and Results

In order to evaluate the HTM and its two modules, experiments in simulated and in realistic environments have been conducted. Simulations allow to modify the complexity of an environment and to focus only on the results of the analyses performed by the computational modules without taking into account any hardware issues. Realistic environments are suitable to assessing performances of artificial systems in the real world, that is classification and localization experts working with data from real objects in real-time, using physical robots with their mechanical limitations and imperfections.

This section briefly presents major results achieved with the HTM, firstly in simulated conditions and, secondly in a testing room where different environments are available. However, before presenting these results, it is necessary to describe what will be evaluated, in both the simulations and in the real world.

3.3.1 The Naive Robot

The HTM model covers several fields of AI and robotic behavior, such as attention, learning, and perception. Moreover, robots endowed with head-movement capabilities are rather rare and, as explained before, there is no *correct way* for a robot to operate—only something that could be qualified as *relevant* as compared to how human beings would behave. Thus, it was necessary to find a reference system to assess whether an HTM-driven robot exhibits a “better” behavior than other systems. In the current study, a “naive” robot \mathfrak{R}_n was employed for this purpose—also referred to as *naive system*. It is similar to the system that Girard et al. (2002) has used and has the following two main characteristics:

1. The naive robot does not perform any further analysis of the data that it gets from the classification experts than concatenating them, that is, the auditory and visual labels are taken from the experts *as is* without any temporal integration or deeper processing.
2. The naive system triggers head movements whenever there is a new audiovisual source appearing in the environment being explored. This behavior could be comparable to a simpler version of the motivation by saliency or novelty. In fact, every time a new object enters the scene, the naive system will guide the robot to focus on it.

A robot driven by the HTM will thus be compared to this naive robot in terms of the quality of the classification and fusion of audiovisual data by the dedicated experts on the one hand, and the number of head movements triggered during the exploration of several environments on the other hand. Since the HTM is a system that *modulates* the head movements by either triggering or inhibiting them as a result of HTM deployment, a significant improvement of the quality of the data from the

experts and a lower number of movements of the head is expected, while maintaining reasonable behavior in terms of the choices as to which objects in the scene should be focused on.

3.3.2 Simulations

Firstly, the simulations mimicked the behavior of the classification and localization experts. The output of the simulated auditory and visual classification experts was emulated, with a specific error rate per frame included in the data generation process in order to reflect the real behavior of the real classifiers. The simulation tool generates probabilities of an auditory/visual frame to belong to a specific auditory/visual class, in addition to a whole virtual environment the robot explores. Thus, one vector per modality, made of as many components as there are experts implemented, will be rendered at every time step. The simulated environments included different numbers of audiovisual sources which can appear anywhere and at any time for durations unknown to the system—see Fig. 10. Two different general cases have been tested; namely, single-source scenarios with no concurring sound sources and multi-source scenarios but only the results from multisource scenarios are presented here. These simulated environments were populated with three to ten overlapping audiovisual sources. All numbers presented are the result of averaging over five runs for each scenario.

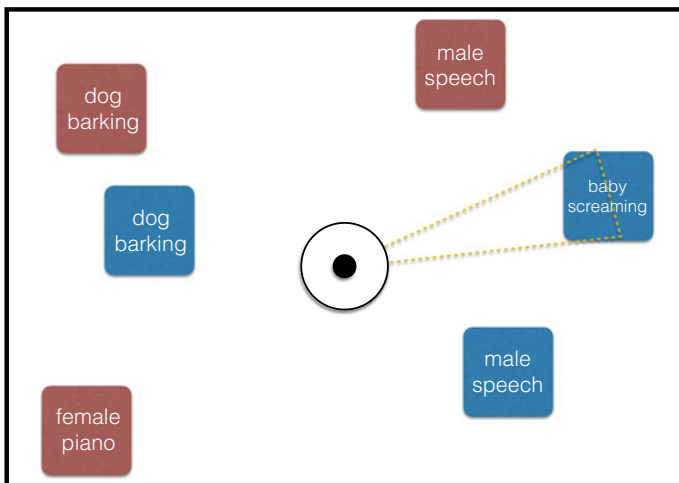


Fig. 10 *Illustration of a simulated environment.* The environments are populated with various audiovisual sources belonging to a certain audiovisual class that the robot does not know beforehand. Some of them are emitting sound (**blue**), others are silent, (**red**). The ability of the robot to acquire “correct” knowledge about the semantic content of the scene is assessed based on congruence of the perceived objects, either via the quality of audiovisual fusion or via the quality of head movements triggering or inhibition

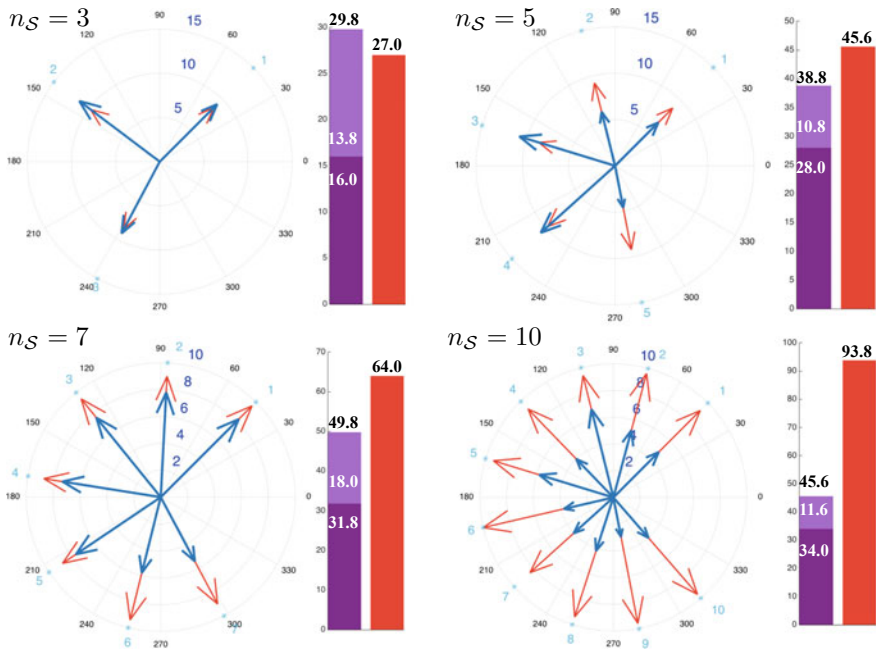


Fig. 11 Number of head movements generated in multisource scenarios. Movements generated by the HTM, (blue), and by the naive robot, (red). The arrows point to the positions of audiovisual sources, their length representing the number of movements toward the considered source. The histograms depict the total sum of generated movements, by the MFI, (dark purple), by the DW, (light purple), and by the naive robot, (red). The (white) numbers correspond to the number of movements by module, averaged over five trials, and their sum (black)

Figure 11 depicts the results obtained under multisource conditions. The histograms are the most interesting data to look at. They depict how many movements were triggered by the DW & MFI modules versus the naive robot. Interestingly, the more complex the environment gets, the more impact the HTM system has on the number of head movements. Indeed, in the scenario with ten audiovisual sources all emitting at the same time, the naive robot triggers up to 93.8 head movements while the HTM triggers only 45.6, that is less than half of them.

3.3.3 Realistic Environments

The experiments performed in realistic environments were conducted with the real robot in a pseudo-anechoic room where several audiovisual sources were placed. The auditory data were emitted by different loudspeakers with QR codes attached to them to identify them as visual objects. Three environments were tested as listed in Table 1. The following audiovisual sources were employed: barking dog, screaming

Table 1 Specification of three scenarios created under real conditions for evaluating the HTM on a real robot

Test-scenario characteristics				
$e^{(i)}$	n_S	n_{sim}^{max}	Present audiovisual classes	Angular position $\theta^{(a v)}$
1	3	1	barking dog #1	320°
			barking dog #2	35°
			speaking male	70°
2	3	1	crying baby #1	70°
			crying baby #2	35°
			piano female	320°
3	3	1	crying baby #1	70°
			crying baby #2	35°
			barking dog	320°
			speaking male	280°

baby, piano female, speaking male. Moreover, the scenarios did not include any whole-body movements because the model addresses head movements only.

Importantly, one of the major roles of the MFI is to *clean up* the data coming from the experts for they exhibit a certain amount of error per frame. To quantify this *clean up* step, a correct audiovisual classification rate $\Gamma(o_j)[t]$ has been set. It is defined by comparing the estimated audio and visual classes (associated to all the sources detected by the system) with the ground truth, according to

$$\Gamma(o_j)[t] = \mathbf{a} \times \sum_{k=t_i}^t \gamma(o_j)[k] \text{ with } \gamma(o_j)[k] = \begin{cases} 1 & \text{if } \widehat{c}(o_j)[k] = c(\Psi_j)[k], \\ 0 & \text{else,} \end{cases} \quad (7)$$

with $c(\Psi_j)[k]$ representing, at time k , the ground truth audiovisual class of event Ψ_j captured as the object o_j in the internal representation of the robot, and $\mathbf{a} = 1/[1, \dots, (t - t_i) + 1]$ as the elapsed time between t_i , the first time step the MFI provided a classification of object o_j , and t the current time. The overall correct classification rate is then given by applying a sliding window on all the $\Gamma(o_j)$ calculated since the exploration has begun, according to

$$\bar{\Gamma}_{\text{MFI}}[t] = \frac{1}{N_{obj}^c[t]} \sum_{j=1}^{N_{obj}^c[t]} \Gamma(o_j)[t], \quad (8)$$

where $N_{obj}^c[t]$ is the number of objects processed so far by the MFI at time t . In parallel, the same process is made for the naive robot \mathfrak{R}_n , along

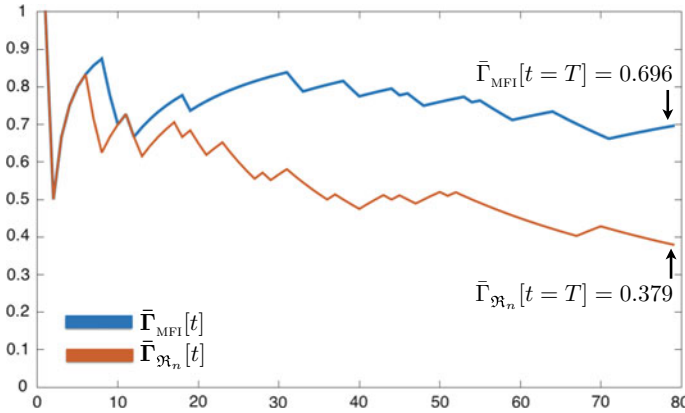


Fig. 12 Average correct audiovisual classification rate computed on a sliding window for the MFI ($\bar{\Gamma}_{\text{MFI}}$), (blue), and the naive robot ($\bar{\Gamma}_{\mathfrak{R}_n}$), (red), that is directly at the classifiers output. The two numbers denote the final results at the end of exploration

$$\bar{\Gamma}_{\mathfrak{R}_n}[t] = \frac{1}{N_{obj}[t]} \sum_{j=1}^{N_{obj}[t]} \Gamma(o_j)[t]. \tag{9}$$

At the end of the exploration, the MFI provides an improvement of about 183.6% in terms of correct audiovisual classification rate, raising from 37.9% for $\bar{\Gamma}_{\mathfrak{R}_n}$ to 69.6% for $\bar{\Gamma}_{\text{MFI}}$. Taking only the labels as assigned by the classification experts would lead to the creation of multiple different audiovisual classes—as illustrated by Fig. 13. This figure illustrates how the MFI considerably narrows the ensemble of possible audiovisual classes: from 22 detected by the experts, the MFI converges to only 5, that is a $\approx 78\%$ diminution. If the DW module had worked directly on the expert’s output, the results of congruence analysis would thus be seriously corrupted. The usefulness of the MFI for the DW module and the robot’s internal representation at large is thus convincingly demonstrated (Fig. 12).

3.3.4 Discussion and Conclusions of Sect. 3

The results presented here for simulated and realistic environments show that the HTM can drastically lower the number of head movements toward unpredictable audiovisual sources based on *congruence* and *reduction of uncertainty* as determined by the DW and the MFI modules. Modulating the generation of such head movements is of importance for achieving suitable means of behavior to separate important from unimportant events. These two modules enable mobile robots endowed with human-like audiovisual perception to explore unknown environments and to react quickly and without prior knowledge to incoming audiovisual objects. The “How”, “Where”, and “When” of the objects need to be determined as they appear in the environment.

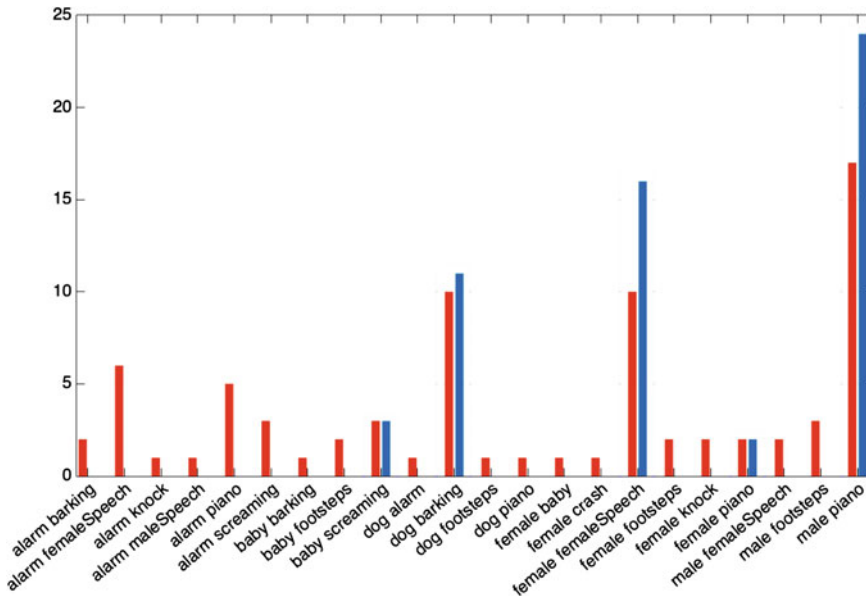


Fig. 13 Audiovisual classes created by the naive robot, (red), and by the MFI, (blue). The height of histograms depicts the number of frames for which the corresponding audiovisual class has been assigned. The **light blue rectangles** highlight the audiovisual classes that are common amongst the two fusion systems

These are first unknown to the system—and thus to the robot. In combination, these two modules form the *Head-Turning-Modulation model* and constitute a complete system, which is working closely together with several experts—classification and localization—in order to establish a form of endogenous attentional behavior in humanoid robots.

4 Final Discussion and Conclusion

Audition and vision are two major senses used by most mammals and humans. Both senses exhibit incredible performances in perceiving and processing the world in their own way. The data that they use are often very complex, be it spatially or temporally, and can change dynamically. The system of very sensitive sensors (eyes and ears) coupled to incredibly powerful means of analysis, such as dedicated sensory areas in the auditory and visual cortical areas, make us understand the real world without too much of an effort. However, when trying to “simulate” such systems, as human-like robotics aim to do, audition and vision are often considered as two separate information channels.

Moreover, it is rather rare to see artificial systems with an additional “cognitive” layer of multimodal integration, allowing the robot to build a deeper internal representation of the world than just a collection of object labels. Also, behavioral rules are often pre-determined by the experimenter leading to “if-else”-statement kind of reactions, such as this one: “*If a baby is crying, go to the baby*”. These kinds of rules might be useful in simple scenarios and for robots with a short lifespan but whenever the robotic agent is put in more complex and varying environments, which have to be explored for extended periods of time (weeks, months, years, ...) the binary priorly defined rules cannot anticipate all the different objects prone to occur. In particular, relevant and comprehensive behavioral rules for properly guiding the exploration will not be available readily.

Thus, the idea of letting the robot create its own behavioral rules was central to the HTM model that is proposed and described here. Inspired by several biological phenomena that are involved in the understanding of the audiovisual perceptual world, the HTM model is an example of how audition can be used as a trigger for head movements towards particular audiovisual sources of interest, thus enabling requisition of data from the visual modality for refining the perception of audiovisual sources of importance. In particular, the results presented in Sect. 3.3.3 provide evidence for the usefulness of multimodal integration of auditory and visual information for a humanoid robot to explore unknown environments when prior knowledge of their audiovisual content is sparse. Moreover, the time needed for the robot to behave adequately and meaningful in unknown environments becomes significantly shorter in this way. Actually, only a few examples are enough for the robot to create its first behavioral rules, thus undermining the widespread misconception that real-time learning and the inability to quickly react in unknown conditions come in couples.

Indeed, the HTM model is far from being the only computational model that integrates several modalities in order to enrich the representation of the world models of robots—see Noda et al. (2014), for instance. However, most current models rely on strong a priori knowledge gained from off-line learning in controlled environments, or on rules available in the form of pre-established “if-else” statements. Such paradigms often prohibit the robots from either learning more from what they experience, or from quickly adapting to situations that have not been encountered before. Yet, the ability to do so is one of the most powerful competencies that human brains have, that is, to quickly adapt to odd situations, be they odd because of their unpredictability or because of their novelty.

Acknowledgements This work has been supported by the European FP7 TWO!EARS project, ICT-618075, www.twoears.eu. We also thank two anonymous reviewers for their previous comments on this work.

References

- Ahissar, M., and S. Hochstein. 2004. The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences* 8 (10): 457–464. <https://doi.org/10.1016/j.tics.2004.08.011>.
- Alain, C., S.R. Arnott, S. Hevenor, S. Graham, and C.L. Grady. 2001. ‘What’, and ‘where’ in the human auditory system. *Proceedings of the National Academy of Sciences of the United States of America* 98 (21): 12301–12306. <https://doi.org/10.1073/pnas.211209098>.
- Alho, K. 1995. Cerebral generators of mismatch negativity (MMN) and its magnetic counterpart (MMNm) elicited by sound changes. *Ear and Hearing* 16 (1): 38–51. <https://doi.org/10.1097/00003446-199502000-00004>.
- Anastasio, T.J., P.E. Patton, and K. Belkacem-Boussaid. 2000. Using Bayes rule to model multi-sensory enhancement in the superior colliculus. *Neural Computation* 12 (5): 1165–1187. <https://doi.org/10.1162/089976600300015547>.
- Arnal, L.H., and A.-L. Giraud. 2012. Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences* 16 (7): 390–398. <https://doi.org/10.1016/j.tics.2012.05.003>.
- Atilgan, H., S.M. Town, K.C. Wood, G.P. Jones, R.K. Maddox, A.K. Lee, and J.K. Bizley. 2018. Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron* 97 (3): 640–655.e4. <https://doi.org/10.1016/j.neuron.2017.12.034>.
- Baranes, A., and P.-Y. Oudeyer. 2009. R-IAC: Robust intrinsically motivated active learning. *IEEE Transactions on Autonomous Mental Development* 1 (3): 155–169.
- Baranes, A., and P.-Y. Oudeyer. 2010. Intrinsically motivated goal exploration for active motor learning in robots: A case study. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IROS, IEEE, 1766–1773. <https://doi.org/10.1109/IROS.2010.5651385>.
- Belin, P., R.J. Zatorre, P. Lafaille, P. Ahad, and B. Pike. 2000. Voice-selective areas in human auditory cortex. *Nature* 403 (6767): 309–312. <https://doi.org/10.1038/35002078>.
- Berlyne, D.E. 1950. Novelty and curiosity as determinants of exploratory behavior. *British Journal of Psychology* 41 (1–2): 68–80.
- Berlyne, D.E. 1954. A theory of human curiosity. *British Journal of Psychology* 45 (3): 180–191.
- Bisley, J.W., and M.E. Goldberg. 2006. Neural correlates of attention and distractibility in the lateral intraparietal area. *Journal of Neurophysiology* 95: 1696–1717. <https://doi.org/10.1152/jn.00848.2005>.
- Blauert, J., and G. Brown. 2020. Reflexive and reflective auditory feedback. In *The Technology of Binaural Understanding*, eds. J. Blauert and J. Braasch, 3–31. Cham, Switzerland: Springer and APA Press.
- Cherry, E.C. 1953. Some experiments upon the recognition of speech with one and two ears. *The Journal of the Acoustical Society of America* 25: 975–979.
- Cherry, E.C., and W.K. Taylor. 1954. Some further experiments upon the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America* 26 (4): 554–559.
- Cohen-L’hyver, B. 2017. Modulation de Mouvements de Tête pour l’Analyse Multimodale d’un Environnement Inconnu [Modulation of head movements for the multimodal analysis of an unknown environment]. Ph.D. thesis, University Pierre and Marie Curie.
- Cohen-L’hyver, B., S. Argentieri, and B. Gas. 2015. Modulating the auditory turn-to reflex on the basis of multimodal feedback loops: The dynamic weighting model. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 1109–1114.
- Cohen-L’hyver, B., S. Argentieri, and B. Gas. 2016. Multimodal fusion and inference using binaural audition and vision. In *International Congress on Acoustics*.
- Cohen-L’hyver, B., S. Argentieri, and B. Gas. 2018. The head turning modulation system: An active multimodal paradigm for intrinsically motivated exploration of unknown environments. *Frontiers in Neurobotics* 12: 60. <https://doi.org/10.3389/fnbot.2018.00060>.
- Corbetta, M., G. Patel, and G.L. Shulman. 2008. Review the reorienting system of the human brain: From environment to theory of mind. 306–324. <https://doi.org/10.1016/j.neuron.2008.04.017>.
- Cuperlier, N., M. Quoy, and P. Gaussier. 2007. Neurobiologically inspired mobile robot navigation and planning. *Frontiers in Neurobotics* 1. <https://doi.org/10.3389/neuro.12>.

- Duangudom, V., and D.V. Anderson 2007. Using auditory saliency to understand complex auditory scenes. In *15th European Signal Processing Conference*.
- Durrant-Whyte, H., and T. Bailey. 2006. Simultaneous localization and mapping (SLAM): Part I. *IEEE Robotics Automation Magazine* 13 (2): 99–110.
- Escera, C., K. Alho, I. Winkler, and R. Naatanen. 1998. Neural mechanisms of involuntary attention. *Journal of Cognitive Neuroscience* 10 (5): 590–604. <https://doi.org/10.1162/089892998562997>.
- Escera, C., E. Yago, M.J. Corral, S. Corbera, and M.I. Nuñez. 2003. Attention capture by auditory significant stimuli: Semantic analysis follows attention switching. *European Journal of Neuroscience* 18 (8): 2408–2412. <https://doi.org/10.1046/j.1460-9568.2003.02937.x>.
- Fendrich, R., and P.M. Corballis. 2001. The temporal cross-capture of audition and vision. *Perception & Psychophysics* 63 (4): 719–725. <https://doi.org/10.3758/BF03194432>.
- Finney, E.M., I. Fine, and K.R. Dobkins. 2001. Visual stimuli activate auditory cortex in the deaf. *Nature Neuroscience* 4 (12): 1171–1173. <https://doi.org/10.1038/nn763>.
- Friston, K. 2005. A theory of cortical responses. *Philosophical Transactions: Biological Sciences* 360 (1456): 815–836. <https://doi.org/10.1080/00222935708693955>.
- Gebhard, J., and G. Mowbray. 1959. On discriminating the rate of visual flicker and auditory flutter. *The American Journal of Psychology* 72 (4): 521–529.
- Girard, B., V. Cuzin, A. Guillot, K.N. Gurney, and T.J. Prescott. 2002. Comparing a brain-inspired robot action selection mechanism with ‘Winner-Takes-All’. In *From Animals to Animats 7: Proceedings of the 7th International Conference on Simulation of Adaptive Behavior*, vol. 7, 75, MIT Press.
- Gurney, K., T.J. Prescott, and P. Redgrave. 2001a. A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics* 84 (6): 401–410.
- Gurney, K., T.J. Prescott, and P. Redgrave. 2001b. A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological Cybernetics* 84 (6): 411–423.
- Hall, J.W., M.P. Haggard, and M.A. Fernandes. 1984. Detection in noise by spectro-temporal pattern analysis. *The Journal of the Acoustical Society of America* 76 (1): 50–56.
- Hay, J.C., H.L. Pick, and K. Ikeda. 1965. Visual capture produced by prism spectacles. *Psychonomic Science* 2 (1–12): 215–216.
- Hochstein, S., and M. Ahissar. 2002. View from the top: Hierarchies and reverse hierarchies review. *Neuron* 36 (3): 791–804.
- Itti, L., C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (11): 1254–1259. <https://doi.org/10.1109/34.730558>.
- Iurilli, G., D. Ghezzi, U. Olcese, G. Lassi, C. Nazzaro, R. Tonini, V. Tucci, F. Benfenati, and P. Medini. 2012. Sound-driven synaptic inhibition in primary visual cortex. *Neuron* 73 (4): 814–828. <https://doi.org/10.1016/j.neuron.2011.12.026>.
- Kayser, C., C.I. Petkov, M. Lippert, and N.K. Logothetis. 2005. Mechanisms for allocating auditory attention: An auditory saliency map. *Current Biology* 15: 1943–1947. <https://doi.org/10.1016/j.cub.2005.09.040>.
- Koch, C., and S. Ullman. 1985. Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology* 4 (4): 219–227.
- Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43 (1): 59–69. <https://doi.org/10.1007/BF00337288>.
- Li, Z. 2002. A saliency map in primary visual cortex. *Trends in Cognitive Sciences* 6 (1): 9–16.
- Lochmann, T., and S. Deneve. 2011. Neural processing as causal inference. *Current Opinion in Neurobiology* 21 (5): 774–781. <https://doi.org/10.1016/j.conb.2011.05.018>.
- Macedo, L., and A. Cardoso. 2001. Modeling forms of surprise in an artificial agent. In *Proceedings of the Cognitive Science Society*, vol. 23.
- Makarenko, A.A., S.B. Williams, F. Bourgault, and H.F. Durrant-Whyte. 2002. An experiment in integrated exploration. In *IEEE International Conference on Robots and Systems*.
- May, P.J. 2006. The mammalian superior colliculus: Laminar structure and connections. *Progress in Brain Research* 321–378. [https://doi.org/10.1016/S0079-6123\(05\)51011-2](https://doi.org/10.1016/S0079-6123(05)51011-2).

- Mazer, J.A., and J.L. Gallant. 2003. Goal-related activity in v4 during free viewing visual search: Evidence for a ventral stream visual salience map. *Neuron* 40: 1241–1250.
- Meredith, M.A., and B.E. Stein. 1986. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology* 56 (3): 640–662. <http://dx.doi.org/citeulike-article-id:844215>.
- Molholm, S., A. Martinez, W. Ritter, D.C. Javitt, and J.J. Foxe. 2005. The neural circuitry of pre-attentive auditory change-detection: An fMRI study of pitch and duration mismatch negativity generators. *Cerebral Cortex* 15 (5): 545–551. <https://doi.org/10.1093/cercor/bhh155>.
- Moschovakis, A.K. 1996. The superior colliculus and eye movement control. *Current Opinion in Neurobiology* 6 (6): 811–816.
- Näätänen, R., and K. Alho. 1995. Generators of electrical and magnetic mismatch responses in humans. *Brain Topography* 7 (4): 315–320.
- Näätänen, R., A. Gaillard, and S. Mäntysalo. 1978. Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica* 42: 313–329.
- Näätänen, R., P. Paavilainen, T. Rinne, and K. Alho. 2007. The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 118 (12): 2544–2590. <https://doi.org/10.1016/j.clinph.2007.04.026>.
- Nahum, M., I. Nelken, and M. Ahissar. 2008. Low-level information and high-level perception: The case of speech in noise. *PLoS Biology* 6 (5): e126. <https://doi.org/10.1371/journal.pbio.0060126>.
- Nelken, I., and M. Ahissar. 2006. High-level and low-level processing in the auditory system: The role of primary auditory cortex. *Dynamic of Speech Production and Perception*: 5–12.
- Noda, K., H. Arie, Y. Suga, and T. Ogata. 2014. Multimodal integration learning of robot behavior using deep neural networks. *Robotics and Autonomous Systems* 62 (6): 721–736. <https://doi.org/10.1016/j.robot.2014.03.003>.
- Nothdurft, H.-C. 2006. Saliency and target selection in visual search. *Visual Cognition* 14 (4–8): 514–542.
- Oliva, A., A. Torralba, M.S. Castelhana, and J.M. Henderson. 2003. Top-down control of visual attention in object detection. In *IEEE International Conference on Image Processing*, vol. 1, 1–4, September 14–17. <https://doi.org/10.1109/ICIP.2003.1246946>.
- Pick, H.L., D.H. Warren, and J.C. Hay. 1969. Sensory conflict in judgments of spatial direction. *Attention, Perception, & Psychophysics* 6 (4): 203–205.
- Posner, M.I., M.J. Nissen, and R.M. Klein. 1976. Visual dominance: An information-processing account of its origins and significance. *Psychological Review* 83 (2): 157–171. <https://doi.org/10.1037/0033-295X.83.2.157>.
- Ruesch, J., M. Lopes, A. Bernardino, J. Hörnstein, J. Santos-Victor, and R. Pfeifer. 2008. Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *Proceedings—IEEE International Conference on Robotics and Automation*, 962–967. <https://doi.org/10.1109/ROBOT.2008.4543329>.
- Saldana, H.M., and L.D. Rosenblum. 1993. Visual influences on auditory pluck and bow judgments. *54* (3): 406–416.
- Scheier, C.R., R. Nijhawan, and S. Shimojo. 1999. Sound alters visual temporal resolution. *Investigative Ophthalmology & Visual Science* 40: S792–S792.
- Schymura, C., Kolossa D. 2020. Blackboard systems for modeling binaural understanding. In *The Technology of Binaural Understanding*, eds. J. Blauert and J. Braasch, 91–111. Cham, Switzerland: Springer and ASA Press.
- Shamma, S. 2008. On the emergence and awareness of auditory objects. *PLoS Biology* 6 (6): e155. <https://doi.org/10.1371/journal.pbio.0060155>.
- Shams, L., C.A.Y. Kamitani, S. Thompson, and S. Shimojo. 2001. Sound alters visual evoked potentials in humans. *Cognitive Neuroscience and Neuropsychology* 12 (17): 3849–3852.
- Shams, L., Y. Kamitani, and S. Shimojo. 2002. Visual illusion induced by sound. *Cognitive Brain Research* 14: 147–152.

- Shams, L., S. Iwaki, A. Chawla, and J. Bhattacharya. 2005. Early modulation of visual cortex by sound: An MEG study. *Neuroscience Letters* 378 (2): 76–81. <https://doi.org/10.1016/j.neulet.2004.12.035>.
- Sharma, J., A. Angelucci, and M. Sur. 2000. Induction of visual orientation modules in auditory cortex. *Nature* 404 (6780): 841–847. <https://doi.org/10.1038/35009043>.
- Spence, C.J., and J. Driver. 1994. Covert spatial orienting in audition: Exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance* 20 (3): 555–574.
- Spence, C., and J. Driver. 1996. Audiovisual links in endogenous covert spatial attention. *Journal of Experimental Psychology: Human Perception and Performance* 22 (4): 1005–1030.
- Spence, C., and J. Driver. 1997a. Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics* 59 (1): 1–22. <https://doi.org/10.3758/BF03206843>.
- Spence, C., and J. Driver. 1997b. On measuring selective attention to an expected sensory modality. *Perception & Psychophysics* 59 (3): 389–403. <https://doi.org/10.3758/BF03211906>.
- Stein, B.E., W. Jiang, and T.R. Stanford. 2004. Multisensory integration in single neurons of the midbrain. *The Handbook of Multisensory Processes*, vol. 15, 243–264.
- Thompson, K.G., and N.P. Bichot. 2005. A visual salience map in the primate frontal eye field. *Progress in Brain Research* 147: 251–262.
- Treisman, A.M., and G. Gelade. 1980. A feature-integration theory of attention. *Cognitive Psychology* 12 (1): 97–136. [https://doi.org/10.1016/0010-0285\(80\)90005-5](https://doi.org/10.1016/0010-0285(80)90005-5).
- Turatto, M., F. Benso, G. Galfano, and C. Umiltà. 2002. Nonspatial attentional shifts between audition and vision. *Journal of Experimental Psychology: Human Perception and Performance* 28 (3): 628–639. <https://doi.org/10.1037//0096-1523.28.3.628>.
- Two!Ears, N. Ma, I. Trowitzsch, Y. Kashef, J. Mohr, K. Obermayer, C. Schymura, D. Kolossa, T. Walther, H. Wierstorf, T. May, G. Brown, B. Cohen-L'hyver, P. Danès, M. Devy, T. Forgue, A. Podlubne, and B. Vandeportaele. 2012. Report on evaluation of the Two!Ears expert system. Technical report.
- Vetter, P., F.W. Smith, and L. Muckli. 2014. Decoding sound and imagery content in early visual cortex. *Current Biology* 24 (11): 1256–1262. <https://doi.org/10.1016/j.cub.2014.04.020>.
- Welch, R.B., and D.H. Warren. 1980. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin* 88 (3): 638.
- Wolfe, J.M. 1994. Guided search 2.0—a revised model of visual search. *Psychonomic Bulletin & Review* 1 (2): 202–238. <https://doi.org/10.3758/BF03200774>.
- Yost, W.A. 1992. Auditory perception and sound source determination. *Current Directions in Psychological Science* 1 (6): 179–184.

Intelligent Hearing Instruments—Trends and Challenges



Eleftheria Georganti, Gilles Courtois, Peter Derleth and Stefan Launer

Abstract Hearing instruments (HIs) aim at helping people with hearing impairment who often have difficulties to understand speech in noisy environments. This chapter provides an overview of the current technological trends and challenges in the field of HI applications. It covers the state-of-the-art of signal-processing algorithms used in modern digital HIs. Focus is given on the extensions of such algorithms for applications, where microphone signals are employed from both the left and right HIs (binaural case). Furthermore, the chapter refers to the challenges for the optimal parametrization and steering of the HI algorithms. The concepts of environment classification for automatically controlling the settings of an HI in different listening situations are discussed and a brief summary of sound-source-localization methods is given. Finally, this chapter discusses the current trends of adding sensors in HIs that can potentially further enhance the hearing performance of the devices and improve the life of hearing-impaired people.

1 Introduction

Hearing impairment is defined as a partial or total inability to hear (Britannica 2017). It may occur in one or both ears and is one of most common physical conditions affecting elderly adults. The World-Health Organization (WHO 2017) reports that over 5% of the world's population (360 million people) has a disabling hearing loss and among them 32 million are children. Hearing loss is nearly always associated with

E. Georganti (✉) · G. Courtois · P. Derleth · S. Launer
Science and Technology, R&D, Sonova AG, Laubisruetistrasse 28, 8712 Staefa, Switzerland
e-mail: eleftheria.georganti@sonova.com

G. Courtois
Signal Processing Laboratory (LTS2), Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

S. Launer
School of Health and Rehabilitation Sciences, University of Queensland, Brisbane, Australia



Fig. 1 Different types of state-of-the-art hearing aids. **a** BTE, **b** RIC, **c** ITE, **d** ITC, **e** CIC

poor speech perception and some areas of greatest difficulty for hearing-impaired include communication in background noise, difficulty to understand talkers with soft voices, hearing speech at a distance, or talking over the phone (Popelka and Moore 2016). Some of these deficits can be addressed with hearing instruments (HIs), as listed below.

- **Hearing aids** are electronic devices that primarily aim at amplifying, filtering and delivering adequate sound signals to the ears. Hearing aids are the most widespread solution to improve the hearing performance of hearing-impaired people. There are different types of them, namely, (a), behind the ear (BTE), (b), receiver in the ear canal (RIC), (c), in the ear (ITE), (d), in the canal (ITC), (e), completely in the canal (CIC)—see Fig. 1.
- **Cochlear implants (CI)** consist of two different units: (i), the speech processor and, (ii), the implant. The speech processor is worn behind the ear and it is composed of a microphone and a microprocessor. The implant includes an electrode array that is implanted in the cochlea and directly stimulates the auditory nerve. The communication between both parts is performed through a wireless connection through the skull.
- Further HIs exist, such as **bone-anchored hearing systems, middle-ear implants and auditory-brainstem implants**, but it is beyond the scope of this chapter to address the specific details of these systems.

Nowadays, all the aforementioned devices include one or more microphone sensors for capturing the audio signal and employ digital signal processing (DSP) methods in order to enhance the input signal accordingly, before delivering it to the listener. The employed signal processing methods are continuously evolving in terms of performance and speed, and they rely on technological advancements in the fields of microelectronics and DSP. This will allow future hearing devices to turn into more intelligent systems offering a range of specific algorithms and algorithmic settings for addressing the specific listening and communication needs of users in different acoustic environments.

This chapter provides a survey of current trends and challenges in the field of HI applications with emphasis on hearing aids and cochlear implants. Section 2 describes state-of-the-art signal-processing algorithms for HIs and discusses current trends regarding the extensions of these algorithms to their binaural versions, that is, employing microphone signals from both the left and right HIs. Section 3

gives an overview of methods for the analysis of the auditory environment, which is essential for optimal parametrization and steering of the HI algorithms. It also describes several recently developed sound-localization methods based on binaural input signals and their extensions when an additional remote microphone is available. Section 4 focuses on the current trends in sensors for HIs and their potential applications. Finally, Sect. 5 summarizes the contents of this chapter.

2 Signal-Processing Algorithms in Hearing Instruments

Over the last decades, the upswing in digital-hearing devices has led to the development of new signal-processing algorithms. The aim of these algorithms is to compensate for the hearing loss of hearing-impaired listeners. Table 1 reports typical signal-processing features as incorporated in HIs and specific issues and hearing deficits that they address (Kollmeier et al. 1993; Courtois 2016; Launer et al. 2016; Souza 2016).

Figure 2 provides an overview of the signal-processing blocks found in HIs—see also Launer et al. (2016). The signals are typically picked up by one or multiple omnidirectional microphones that are placed within the devices. In the case where two devices, left and right, are available and wirelessly connected, the microphone signals from both devices may be utilized—binaural HIs. Moreover, alternate audio-source signals may also be present and can be captured with an external microphone and wirelessly streamed. The signals are then processed accordingly and delivered to the listeners by either loudspeakers, electrical signals (cochlear implants), or vibrations (bone-anchored hearing systems). The subsequent algorithms can be broadly classified, according to their delivered functionality as follows, (i), environment analysis, (ii), sound “cleaning”, (iii), audibility and loudness restoration. For general information regarding typically employed DSP algorithms in HIs, see, for example, Kates (2008), Launer et al. (2016), Holube and Pudder (2014), Dillon (2012), further Hamacher et al. (2005, 2006) and Edwards (2007).

Table 1 Typical signal processing features available in HIs and the deficits they address

DSP features	Addressed deficits related to hearing impairment
Amplification	Decreased audibility of sounds
Compression	Abnormal perception of loudness (recruitment)
Frequency lowering	Loss of high-frequency audibility
Beamforming	Deteriorated speech understanding in complex conditions
Noise reduction	Listening discomfort in noisy situations
Dereverberation	Listening discomfort in reverberant environments
Feedback cancellation	Listening discomfort related to the occurrence of feedback

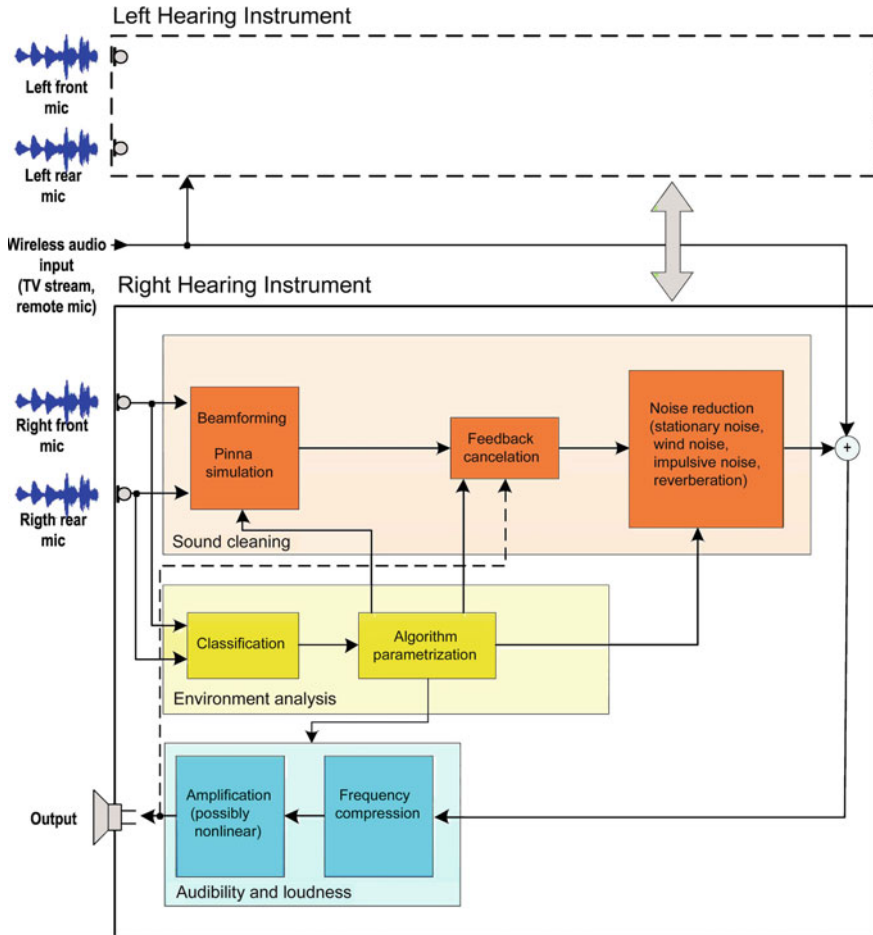


Fig. 2 Generic signal-processing scheme, as typically incorporated in modern HIs equipped with two microphones (**front and rear**). The subsequent algorithms can be classified in three categories, namely, sound “cleaning” (**orange box**), environment analysis (**yellow box**), and audibility and loudness restoration (**blue box**). An additional wireless-audio input is displayed, providing a signal, for instance, from a remote microphone or a TV stream. Figure based on Launer et al. (2016)

Over the past few years, there has been a growing interest in the extension of widely used DSP features in the so-called binaural HIs. These are devices that are capable of exchanging data ear-to-ear (e.g., by a wireless link). This section discusses these aspects and gives an overview of the potential improvements that they could bring. Section 2.1 provides some general information with respect to binaural HIs, and Sects. 2.2–2.4 refer to binaural extensions of typically used DSP algorithms, such as beamforming, wide dynamic-range compression (WDRC), noise reduction, and dereverberation.

2.1 *From Bilateral to Binaural Processing*

The concept of binaural HIs was introduced in the 1990s, when the idea of transmitting the audio signal from one device to the other emerged (Kollmeier et al. 1993; Kollmeier and Koch 1994; Wittkop et al. 1996). The main objective was to estimate the binaural cues, in order to discriminate between the sub-bands dominated by the target speaker and the sub-bands dominated by the undesired components (e.g., competing speakers, diffuse noise, late reverberation). Assuming that the position of the speaker of interest was known, the strategy was to let those frequency bands pass that yield interaural time differences (ITD) and interaural level differences (ILD) that correspond to the direction of arrival (DOA) of the desired speech. Simultaneously, the other bands are attenuated because they presumably are degraded by noise and reverberation (Kollmeier et al. 1993).

It is common to distinguish between two types of binaural HIs, that is, the ones that provide a synchronization-based processing, and the ones that offer a streaming-based approach (Moore 2007a). The first type denotes HIs that share local features parameters between the two devices, such as volume-control levels, program selections or WDRC settings. The second type refers to HIs that exchange complete audio streams. It has been shown that synchronization-based HIs can preserve the original binaural cues, and thus improve sound-localization performance of the listener. However, they fail to enhance speech understanding. Also, they are not always judged by HI users as being more attractive than unlinked devices (Smith et al. 2008; Sockalingam et al. 2009; Ibrahim et al. 2013). The second category of binaural instruments, allowing for full or partial audio streaming between both devices, can potentially lead to considerable improvements in the performance of certain signal-processing algorithms as implemented in HIs, such as beamforming, noise cancelling (Timmer 2013). These algorithms will be discussed in the next sections.

2.2 *Beamforming*

HIs often have at least two omni-directional microphones; typically one located on top of the device (front microphone) and one behind (back microphone), both looking to the front. The most basic process of beamforming can be achieved by delaying the signal from one microphone and adding or subtracting the outputs of the two microphones. With this approach, one can create a “beam pattern” that points to a specific, usually frontal, direction. In this way, target signals from the direction of the beam are picked up well, while sounds from the sides or rear are attenuated. By varying how the outputs of the two microphones are delayed and combined, different beam patterns can be generated (Soede et al. 1993; Stadler and Rabinowitz 1993; Widrow and Luo 2003; Elko and Meyer 2008). Beamforming can be performed independently in the left and right HIs (bilateral beamformer) or the microphone signals from both devices may be used in an appropriate combination—binaural

beamformer. Here it should be stated that beamforming leads to signal-to-noise ratio (SNR) improvements, however some distortions of the spatial cues and a reduction of the overall speech quality may occur (Keidser et al. 2006, 2009; Van den Bogaert et al. 2011).

Advanced designs of beamforming include the minimum-variance distortionless response (MVDR; Capon 1969), the linearly constrained minimum variance (LCMV; Frost 1972), the multi-channel Wiener filter (MWF; Widrow et al. 1975), and the generalized sidelobe canceler (GSC; Griffiths and Jim 1982). The MVDR constrains the beamformer response to capture the sound source in the desired look direction without distortion, while attenuating signals in the other directions. The LCMV is similar to a MVDR, but employs different constraints (linear expressions). The MWF uses the second-order statistical properties of the signal to estimate the desired speech component in a minimum mean-square-error (MMSE) sense in one of the received microphones. Several extensions of the MWF have been proposed, such as the speech-distortion weighted MWF (SDW-MWF) that introduces a trade-off coefficient between noise reduction and speech distortion (Doclo and Moonen 2002), or the MWF with partial-noise estimation (MWF-N) (Klasen et al. 2007), in which the constraint of spatial-cue preservation is introduced. Finally, the GSC is an extension of the MVDR in which the constraints are split into two simultaneous and orthogonal operations. It is beyond the scope of this chapter to detail the various algorithms that have been proposed to achieve bilateral beamforming so far—see Doclo et al. (2010) for a thorough review.

The introduction of streaming capabilities between the left and right hearing devices led to the extension of the aforementioned bilateral beamforming approaches (MVDR, LCMV, MWF) to their binaural versions. From HIs embedding N microphones, the ear-to-ear communication gave access to a network of $2N$ microphones—see Fig. 3. This enabled the development of optimal tuning approaches for the various beamforming algorithms, so that they take into account the trade-off between the maximization of the directivity and the preservation of the binaural cues (Widrow and Luo 2003). The preservation of the cues has been addressed by, (i), inserting additional constraints to preserve the ITD in the GSC beamformer (Desloet et al.

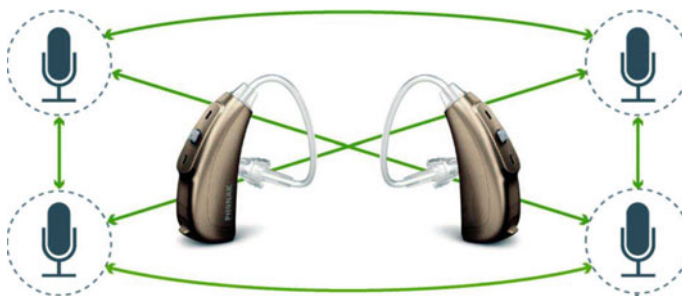


Fig. 3 Two binaural HIs with two microphones each forming a four-microphone network. Reproduced courtesy of Timmer (2013)

1997), (ii), performing beamforming in the high frequencies only and leaving the low-frequency binaural cues unchanged (Welker et al. 1997), (iii), ensuring an accurate reproduction of the ITD and ILD in the frontal areas (Nishimura et al. 2002), (iv), introducing the binaural MWF (BMWF) beamformer that was shown to preserve the ITD of the targeted speech cues (Klasen et al. 2005; Doclo et al. 2006), and (v), introducing the binaural MVDR (BMVDR) beamformer (Doclo et al. 2010).

Research activities in the field of binaural beamforming are continuously growing since the improvement of speech intelligibility in complex (noisy) acoustic environments is one of the primary challenges for HIs. Latzel (2013) describes a binaural beamformer that first computes a standard monaural beamformer at both sides, then processes the beamformed outputs at the ipsilateral device with a predefined weighting function. Hadad et al. (2012) and Marquardt et al. (2014) introduced the binaural LCMV (BLCMV) beamformer with additional constraints related to the preservation of the interaural transfer function (ITF), the magnitude and phase of which represent the ILD and ITD respectively. The interaural cues of the target and interference sources are preserved, but the cues associated with the stationary noise are transposed to the ones of the target (Hadad et al. 2016). Cornelis et al. (2014) proposed an extension of the binaural SDW-MWF and MWF-N beamformers in which a combination of a spatial filter and a common spectral post-filter is applied to both devices. In this way, the speech ILD distortions are suppressed without reducing the noise reduction performance. Their algorithms were designed to work with a reduced bandwidth of the ear-to-ear wireless link, that is, without exchanging every microphone signal available in each hearing device. Liao et al. (2015b, a) revisited the BMWF beamformer, incorporating the a-priori knowledge of a database of acoustic transfer functions (ATFs) measured in an anechoic chamber for various known DOAs. ATFs encapsulate the head-related transfer function (HRTF), the transducer-related, and room-related effects. The technical constraints related to HIs (computational cost, real-time framework, power consumption) were also taken into account in the design of the algorithms.

While several solutions have been reported to preserve the spatial cues of the targeted speech, the question of distortions of the binaural cues related to undesired noise remains. Marquardt et al. (2015) addressed this issue and introduced an extension of the BMWF that includes a criterion for the preservation of the interaural coherence (IC) of the noise components in a diffuse noise field. Thiemann et al. (2016) proposed to preserve the spatial information from those noise in the time-frequency units where the SNR is low. They coupled the binaural MVDR beamformer with an SNR-dependent binary classifier that states whether the signal must be processed by the binaural beamformer (when speech is dominant) or be attenuated (when noise is dominant). With the same objective of preserving the binaural noise cues, Szurley et al. (2016) investigated the binaural MWF-N beamformer and used the contribution of an additional remote microphone worn by the speaker of interest, a configuration typically encountered with the use of frequency modulation (FM) systems—see Sect. 4.1. It has theoretically been proven that the integration of this high-SNR signal in the beamformer increases the output SNR and better preserves the noise-related spatial cues (ITD and ILD). This was indeed confirmed in a simulation. The use of

Table 2 From bilateral to binaural beamformer algorithms. MVDR: Minimum Variance Distortionless Response; LCMV: Linearly Constrained Minimum Variance; MWF: Multi-channel Wiener Filter; GSC: Generalized Sidelobe Canceler; BMVDR: Binaural MVDR; BLCMV: Binaural LCMV; BMWF: Binaural MWF

Description	Beamformers			
Bilateral beamformers	MVDR Capon (1969)	LCMV Frost (1972)	MWF Widrow et al. (1975)	GSC Griffiths and Jim (1982)
Binaural beamformers	BMVDR Doclo et al. (2010)	BLCMV Hadad et al. (2012)	BMWF Klasen et al. (2005)	
Improvements and preservation of the binaural cues of speech		Marquardt et al. (2014)	Cornelis et al. (2014) and Liao et al. (2015a)	
Improvements and preservation of the binaural cues of speech and noise	Thiemann et al. (2016)		Marquardt et al. (2015) and Szurley et al. (2016)	

an external microphone can also be found in the study by Yee et al. (2017), where ITE HIs are considered, that is, devices that incorporate only one microphone. Binaural beamforming can be performed with an ear-to-ear wireless link, as previously reported by Srinivasan et al. (2008), but, due to the symmetrical arrangement and the inherent front/back ambiguity, it is restricted to lateral noise reduction. The authors in the aforementioned publication show that the presence of an additional microphone, located at a distance of 30cm from the HI user and in an optimal azimuth range between 10° and 30° in front, can yield significant improvement of the SNR for both stationary and non-stationary noise sources positioned at 180° in the rear.

Ongoing research activities regarding binaural beamforming, combining strong noise reduction, limited speech distortion, and spatial-hearing preservation, have been reviewed in this section. The evolution of the various bilateral to binaural beamformer approaches is summarized in Table 2. It is evident that the challenges related to the determination of an optimal trade-off between noise reduction and speech quality, and the preservation of the binaural cues, still remain. Other aspects, such as computational complexity and continuous ear-to-ear communication, should also be taken into account.

Clinical Investigation

Over the past few years, numerous clinical investigations in order to determine the benefits provided by the use of binaural against bilateral beamformers and the determinant factors supporting both approaches, have been conducted. Picou et al. (2014) investigated the effect of three types of beamformers (mild, moderate, or strong) in eighteen hearing-impaired listeners with regard to speech intelligibility, subjec-

tive preference, and listening effort. The first beamformer in this test provided fixed bilateral directional filtering in high frequencies only. The second one was a bilateral adaptive beamformer, where the maximum attenuation occurred at the main location of the noise sources, and the third was a cue-preserving binaural beamformer. It was evidenced that cue-preserving processing significantly enhanced the speech intelligibility in a realistic listening scenario as compared to the two other types—including reverberation and diffuse background noise. However, there was no significant difference in subjective preference and listening effort between the three beamformers. Appleton and König (2014) compared four beamformer settings, namely, omnidirectional (i.e. no beamformer), bilateral adaptive beamformer, binaural static beamformer, and binaural adaptive beamformer. By testing twenty hearing-impaired listeners and presenting noise either diffused or located on the sides only, they showed that the binaural beamformers outperformed the bilateral one for speech understanding as well as in subjective rating tasks for both noise presentations. When the noise was played back on the sides only, the results revealed better performance of the binaural adaptive beamformer when compared to the static one in both objective and subjective outcomes. Froehlich et al. (2015) conducted a multicentric clinical study on 29 normal-hearing and 43 hearing-impaired listeners in order to compare speech-intelligibility performance obtained with a bilateral omnidirectional directivity pattern and a binaural beamformer. A decrease of the speech-reception threshold (SRT) by 5 dB was observed in the hearing-impaired listeners when using the binaural beamformer in a configuration with diffused babble noise. The speech-intelligibility gain was more pronounced in the hearing-impaired group than in the normal-hearing group.

In a comprehensive study aiming at finding correlations between the individual factors of 60 hearing-impaired listeners (hearing loss, noise sensitivity, personality etc.) and their preferences for signal-processing features (including beamforming), Neher et al. (2016), noticed that the tested binaural beamformer was more appreciated by listeners with a high degree of hearing loss compared to the ones having a moderate hearing impairment. It also appeared that the binaural approach was preferred in a single-talker paradigm (speaker located at 0°), whereas the bilateral beamformer was more attractive in a two-talker scenario (speakers at $\pm 30^\circ$). In a further clinical research, Neher et al. (2017) exposed 39 hearing-disabled listeners to different beamformer settings that represented various trade-offs between SNR improvement and binaural cue preservation. They found that subjects exhibiting a binaural intelligibility level difference (BILD) higher than 2 dB benefited more from the preservation of the low-frequency binaural cues (< 800 Hz) in terms of speech intelligibility, despite the smaller SNR improvements. The opposite was observed in the listeners presenting lower BILD, where the spatial-hearing preservation was of less importance. Geetha et al. (2017) tested various configurations of HI fittings, including beamforming off or on. Speech intelligibility in noise and sound-localization experiments were conducted on a panel of 25 hearing-impaired subjects. The authors found a significant improvement in speech understanding with beamforming, whatever the DOA of the babble noise.

The present review shows that binaural beamformers yield significant enhancements of speech intelligibility in noisy situations. However, clinical studies have also pointed out the strong influence of individual factors and acoustic configurations on the efficiency of such beamformers. Here it should be noted that one of the most dominating factors in practical applications affecting the performance of beamformers is the acoustic coupling to the ear. This refers to the fact that various types of fittings are used when fitting an HI, mainly depending on the type of hearing loss (open domes, closed domes, ear tips etc.). The type of fitting together with the hearing loss of the HI user in the low frequencies determines the direct sound dominance and the resulting listening benefit for the HI user. Therefore, it is of great importance to always take into account the variability of acoustic couplings to the ear before extracting general conclusions about the performance of specific beamforming modes.

2.3 Noise Reduction

Noise reduction systems incorporated in HIs are primarily intended to increase the comfort and diminish the listening effort of the listener in noisy environments. The majority of current bilateral noise cancellers do not provide substantial SNR improvement, and thus fail to improve speech intelligibility—see Moore (2007b) for a review. For the past few years, binaural processing has been introduced in noise reduction algorithms and has shown to provide promising performance. Yang et al. (2013) suggested to compute the ITD and IC between both devices in the low frequencies and the ILD and IC in the high frequencies. Assuming a target speech at 0° , the estimated cues drive a binary classifier that discriminates between the time-frequency regions dominated by the desired speech (corresponding to low ILD/ITD values and high IC values) and the regions where the noise is dominant. Yousefian et al. (2014) reported a noise-reduction algorithm that relies on the calculation of the IC between both devices. A coherence-based gain function is applied similarly on the left and right signals so that the time-frequency components characterized by a low (estimated) SNR are attenuated, while the binaural cues are left unchanged. The algorithm was assessed with coherent noise sources (competing speaker or speech-shaped noise), spatially separated from the target speech located at 0° . An average SRT increase by 6.5 dB was found in eight normal-hearing listeners, but this improvement tends to diminish with increasing reverberation time.

Another common noise type present in HI application is the wind noise, which is mainly caused by the airflow around the head. Hiruma et al. (2016) addressed this issue and proposed a binaural approach for the cancellation of wind noise. Since such a type of noise has typically low-frequency components, the authors implemented a frequency-warping technique to ensure high resolution in the low frequencies and a limited computational delay. The algorithm assumes that the DOA of the target speech is known, and computes the frequency-dependent error between the expected binaural cues obtained from a HRTF database and the real-time estimated cues. That error constitutes a wind-occurrence detector and monitors the wind noise canceller.

2.4 *Dereverberation*

Dereverberation algorithms are used to reduce the undesired effects of reverberation. In general, approaches similar to noise reduction can be adopted for dereverberation. An overview of binaural dereverberation algorithms can be found in Tsilfidis et al. (2013). In the dereverberation algorithm proposed by Westermann et al. (2013), the short-term IC is computed between the left and right HIs microphone signals. A non-linear sigmoid mapping associates some gain values to the computed coherence. The parameters of this mapping are updated online, based on the time-frequency behavior of the IC. Ten listeners went through a subjective assessment of the processing, which allowed to determine the best parameters and the optimal range of speaker-to-listener distances for which the algorithm provides the best results. Marquardt et al. (2013) and Braun et al. (2014) proposed an application of the BMWF including the preservation of the IC of a stationary diffuse noise field (BMWF-IC) to process dereverberation in binaural hearing devices. They considered a time-varying diffuse sound field and resorted to a spherical model of the head to determine the optimal Wiener filter. Schwartz et al. (2015) addressed the dereverberation by introducing a recursive expectation-maximization algorithm. Their goal was to develop a dereverberation method that offers a direct control of the trade-off between dereverberation performance and ITF preservation, which they theoretically demonstrated. With the implementation of a Kalman filter, their algorithm estimates the desired early signal and the short-term room impulse responses modeled by an exponential decay.

2.5 *Wide Dynamic-Range Compression*

Most people with sensorineural hearing loss experience loudness recruitment, that is, once the level of a sound exceeds the elevated absolute threshold, the loudness grows more rapidly than normal with increasing sound level (Fowler, 1936; Moore 2007a, b). In general, the greater the hearing loss, the greater is the rate of the growth of loudness (Miskolczy-Fodor, 1960). However, individual variabilities can be considerable. Typically, HIs process sounds in 5–20 frequency channels, whereby the bandwidth of the channels increases with increasing center frequency. In each of these channels, a level-dependent gain is applied. To compensate for loudness recruitment, the gain should decrease progressively with increasing input level, meaning that the input-output function is compressive. This function is called wide dynamic-range compression, WDRC (Killion 1979), and is widely used in HIs. WDRC may be applied either independently in the left and right HIs (unlinked) or by using the same compression settings on both sides (linked).

Wiggins and Seeber (2012) investigated the effect of unlinked fast-acting WDRC on spatial perception. WDRC is known to reduce the range of ILD and introduces undesired fluctuations of this cue when operating independently in both devices. This yields conflicts between the (unaffected) ITD and the distorted ILD, which

can impair sound localization. Eleven normal-hearing listeners took part in their study and rated various spatial attributes. It was shown that unsynchronized WDRC might increase diffuseness, create potential image splits, and reduce externalization of sound sources. Schwartz and Shinn-Cunningham (2013) compared the effect of independent and linked compressions on 39 normal-hearing subjects. With the unlinked fast-acting compression, the listeners needed a higher spatial separation between target and masker to maintain their spatial selective auditory attention than with synchronized compression. This could be an indication that independent compression might potentially demand a stronger listening effort in noisy situations. However, this observation requires further investigations.

Korhonen et al. (2015) evaluated the effect of linked compression on the localization performance of ten hearing-impaired listeners in the horizontal plane. In this configuration, the gain at the two ears is set to the gain computed at the side with the louder sound source. Although the linked compression was found to preserve the ILD cues, its positive effect on the localization abilities of subjects was rather small and not statistically significant. Hassager et al. (2017b) tested twelve normal-hearing and twelve hearing-impaired listeners in an experiment where they were asked to describe their spatial impression (position and width of sound images) for three types of WDRC, namely, independent compression, linked compression, and spatially ideal compression. In the latter case, the WDRC was applied first, and then the stimuli were spatialized with a pair of head-related impulse response (ideal configuration). It was found that both independent and linked fast-acting compressions yielded wider sound images and could cause internalization and image splits in both normal-hearing and hearing-impaired listeners. Only the ideal compression preserved the spatial impression found with linear amplification. Measurements of the ITD, ILD, the interaural coherence (IC), and direct-to-reverberant ratio (DRR) indicated that IC and DRR were distorted with the linked compression. These results evidence that the preservation of the ILD with synchronized compression is not sufficient to guarantee an accurate spatial perception. In a further publication, Hassager et al. (2017a) proposed a compression scheme based on a binary classifier that determines the time segments dominated by direct sound or by reverberation. When direct sound is dominant, the linked compression is left unchanged, while in areas dominated by reverberation, the gain model is linearized in order to avoid excessive amplification of the sound reflections. This processing was shown to better preserve the IC than conventional linked compression. Also, subjective ratings on 18 normal-hearing listeners indicated that the resulting spatial experience was similar to the one obtained with linear amplification. More details about this approach can be found in May et al. (2020), this volume.

2.6 Summary

Sections 2.3–2.5 have shown that most of the DSP algorithms currently implemented in HIs may take advantage of binaural extensions, in order to improve the spatial

Table 3 Binaural DSP algorithms for HI applications

DSP features	Studies	Method	Objectives
Dynamic-range compression	Schwartz and Shinn-Cunningham (2013)	Linked Compression	Improve spatial perception and reduce listening effort
	Korhonen et al. (2015)	Linked compression	Improve sound localization
	Hassager et al. (2017a)	Direct-sound driven linked compression	Preserve spatial perception
Noise reduction	Yang et al. (2013)	ITD, ILD, and IC-based speech/noise classifier	Improve speech intelligibility and reduce speech distortion
	Yousefian et al. (2014)	IC-based gain mapping	Improve speech intelligibility
	Hiruma et al. (2016)	IC-based gain mapping (wind noise application)	Improve speech intelligibility and spatial perception
Dereverberation	Westermann et al. (2013)	IC-based gain mapping	Improve speech intelligibility
	Braun et al. (2014)	BMWF including IC preservation of the noise	Improve speech intelligibility and spatial perception
	Schwartz et al. (2015)	Expectation-maximization method Kalman filter implementation	Improve speech intelligibility and spatial perception

experience and increase their beneficial effect. Table 3 provides a summary of the binaural realizations that have been reviewed for dynamic-range compression, noise reduction, and dereverberation. A promising outcome is that many algorithms that were previously known to increase solely the listening comfort may also be able to significantly enhance speech understanding in complex acoustic environments.

3 Auditory-Environment Analysis

The analysis of the auditory environment surrounding the HI user is essential for the optimal parametrization and steering of the DSP algorithms that were discussed in Sect. 2. Here, an overview of methods for the analysis of the auditory environment is provided. Section 3.1 discusses standard approaches implemented in modern HIs for the classification of the auditory scene. Section 3.2 describes state-of-the art sound-localization methods that take into advantage the wireless link between the left and right hearing devices and, potentially, an additional remote microphone.

3.1 Auditory Scene Classification

HIs contain a broad range of signal processing strategies designed to provide good speech-intelligibility performance in different listening and acoustic conditions. Algorithms such as noise reduction or acoustic feedback-management systems should be switched on and off or tuned accordingly to deal with different listening situations—see Sect. 2. A classic example is when the HI user moves from a quiet listening environment to a noisy restaurant. In the quiet condition, the microphone characteristic could be omni-directional to allow the detection of sounds from all directions, while in the noisy restaurant a directional microphone characteristic would provide better speech intelligibility. Listening to music might require settings different from those for listening to speech—and the optimal music settings might even depend on the type of music. The monitoring of the acoustic environment can also be used to tune the parameters of the compression and the amplification in real time, as detailed in May et al. (2020), this volume.

In order to tackle the variation of the listening environments and to tune accordingly the various algorithms, many modern HIs contain multiple *scene-classification “programs”* to be set up for different listening situations. Current programs incorporated in many HIs are, for example, calm situation, speech in noise, speech in loud noise, speech in car, music. These programs are customized for specific listening environments. Switching from one to another causes alterations of the time constants that control the speed of the compressor, the amplification and compression settings, the noise-reduction parametrization or the beamforming settings. These programs can be selected manually, such as by pressing a button on the device, or switched from one to another automatically. The need for an automatic program selection is evident, given the fact that many HIs users may be unsuccessful at changing the manual programs appropriately (Ricketts et al. 2017). For this reason, some automatic *“environment-control” algorithms* have been introduced into modern HIs. These algorithms are based on the extraction of various acoustic features for classifying a listening environment by comparing the observed values of the features with a pre-stored map of values (Kates 1995; Nordqvist and Leijon 2004; Büchler et al. 2005). Such systems have their roots in the early work of Bregman (1990).

In HI applications, typically supervised machine-learning approaches are employed, where the system is trained using labeled examples of sounds from each of target-sound class (Bishop 2006). In Fig. 4, a schematic block diagram of an automatic classification algorithm is shown. During the development stage of such algorithms, the system takes as an input audio recordings that are labeled according to the class that they belong to, for example, speech in quiet, music, or speech in noise. Then, several acoustic features are computed and acoustic models are extracted based on machine-learning approaches—such as hidden Markov models, maximum likelihood, Bayesian estimators or neural networks. Typically, features like sound level, spectral centroid, spectral flux, short-time energy, level differences across different frequency bands, or spectral shapes are used—see also Chap. 4 in Popelka and

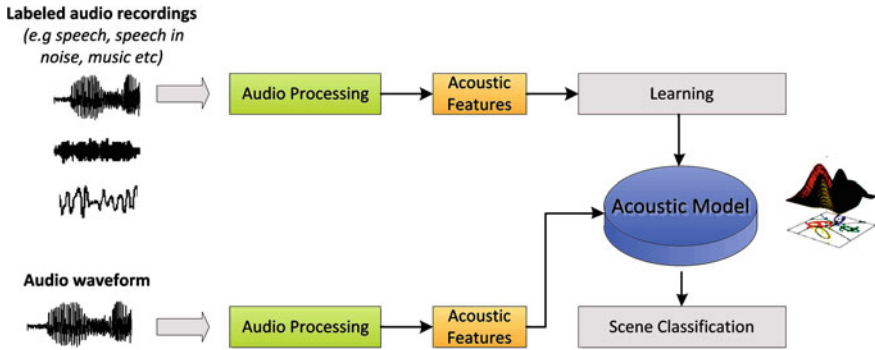


Fig. 4 Schematic block diagram of an automatic classification algorithm for auditory environment classification

Moore (2016). Based on these features, acoustic models are extracted for each class and the auditory scene is classified accordingly.

3.2 Localization of Sound Sources

For many HI applications, it could be useful to know not only the class of the auditory scene, for instance speech, in quiet, music, or speech in noise, but also the position(s) of the sound source(s) relative to the listener. This allows for “smarter” signal-processing and enhancement schemes, for example, for automatic adjustment of the look direction of the beamformer towards the target source location.

In the last decades, a new research field termed computational auditory scene analysis (CASA) has emerged that aims at reproducing the capabilities of the human auditory system with machines on the basis of sensory input (Wang and Brown 2007; Ellis et al. 2018). These methods aim at achieving human performance as regards sound-source localization, recognition, and separation, by using one or two microphone recordings of the acoustic scenes. Inspired by the robustness of the human auditory system, several studies have incorporated stages of human auditory processing to improve sound-source localization in adverse acoustic conditions (Bodden 1993; Faller and Merimaa 2004; Wilson and Darrell 2005). A comprehensive overview of CASA and its relevance for applications in the field of automatic speaker recognition (ASR), speaker localization, and speech segregation can be found in Wang and Brown (2007) and Courtois et al. (2014). Some other examples of recently developed methods may be found in Merks et al. (2013), May et al. (2013), Courtois et al. (2015b) and Anemüller and Kayser (2017). It is beyond the scope of this chapter to provide an overview of all existing sound-source localization methods. The interested reader is referred to the aforementioned studies—for a more general view on algorithms

that rest on biological paradigms compare, for example, Blauert and Brown (2020), this volume.

In this section, emphasis is put on recent methods that were developed and take the limitations of HIs into account, such as constraints in terms of computational capability and memory. HI processors have to work at low clock rates to minimize the power consumption and thus maximize the battery life. Additionally, the restrictions become stronger since a considerable part of the computational capabilities of the processor is already being used for the audio processing aiming to compensate for hearing loss in an “environment-specific” way, as reported in Sect. 2. When it comes to sound-source localization, another challenge is that such algorithms should act very fast, that is, in a few milliseconds, and take account head movements and should be able to recognize the auditory environment even when the sound sources are moving in complex auditory scenes.

In the following, some examples of recently developed localization methods for HIs are given. Braun et al. (2015) proposed an algorithm that relies on the online estimation of the direct-sound-relative transfer function (RTF) between both HI microphones. This estimate is compared to a database of reference RTFs, measured in anechoic conditions, to find the azimuth angle of the present sound source. The localization is performed in the full horizontal plane via the use of HIs having multiple microphones. Courtois et al. (2015b) developed a binaural localization algorithm that computes the short-term interaural phase differences (IPD) and compares them with reference IPD values derived from a spherical model of the head for various azimuths. The algorithm assumes the presence of an additional remote microphone worn by the speaker. This configuration is typically encountered in the context of FM systems. The IPD-based localization is combined with the calculation of the ILD and the *received-signal-strength-indication difference* (RSSID), to determine the position of the speaker in the frontal horizontal plane. The localization algorithm of Farmani et al. (2015) is further based on the accessibility of a clean version of the target signal, as delivered by a remote microphone. This clean signal enables online estimation of the HRTFs at the microphones of a pair of binaural HIs. The inferred DOA is found by looking for the most resembling associated (reference) HRTF in a maximum-likelihood sense. A database of anechoic HRTFs is therefore required. This procedure is repeated for each microphone of the HIs, which increases the reliability of the algorithm results. Zohourian and Martin (2016) addressed the topic of binaural localization by proposing a MMSE-based approach that uses a joint ITD and ILD model. The current observations of the ITD and ILD are computed in each time-frequency unit, and the DOA is the azimuth angle that minimizes the error between the theoretical binaural-cue values and the observation. The localization algorithm is then used to control an adaptive GSC beamformer.

The aforementioned methods are summarized in Table 4. The approaches suggested by Courtois et al. (2014) and Farmani et al. (2015) allow to avoid ear-to-ear audio streaming, however they require an extra remote microphone. In this context, the ability to localize the microphone wearer in real-time offers interesting possibilities for more effective parametrization of the HI algorithm as discussed in Sect. 4.1.

Table 4 Reported binaural localization algorithms for HIs. RTF: Relative Transfer Function; RSSID: Received Signal Strength Indication Difference

	Binaural localization algorithms			
Authors	Braun et al. (2015)	Courtois et al. (2015b)	Farmani et al. (2015)	Zohourian and Martin (2016)
Extracted features	Direct sound RTF and coherent-to-diffuse ratio	IPD, ILD, and RSSID	HRTF	ITD and ILD
Reference	Anechoic RTFs	Spherical model of head	Anechoic HRTFs	HRTF model
Extra microphones	No	Yes	Yes	No
Binaural streaming	Yes	No	No	Yes

3.3 Prognostication

In the future, the fast evolving CASA methods could potentially serve as the basis for the development of novel features for improving the performance of hearing devices. Keeping in mind the continuous evolution of microelectronics with regard to, for example, processing speed, memory capacity, sensors, and size (Mollick 2006). The interest in approaches based on neural networks is growing for various HI functionalities, such as sound-source localization (Ma et al. 2017) or noise and reverberation reduction (May 2018). This, together with the continuous efforts to better understand the importance of visual cues (Varghese et al. 2012; Moradi et al. 2017; Wu and Bentler 2010), human cognitive functions in fields such as selective auditory attention (Shinn-Cunningham and Best 2008) and the introduction of novel intelligent sensors for tracking cognitive functions (Lorenz et al. 2017), could open up a new era in the field of HI processing and performance.

4 Sensors in Hearing Instruments—Present and Future

In this section, an overview of the current trends with respect to sensors and HIs is given. Most of the current HIs rely on a number of microphones placed on the hearing devices to capture the acoustic signal. A typical pair of HIs consists of four microphones, with two microphones on each side. However, several products exist on the market that incorporate a higher or lower number of microphones. Remote microphones are also often used in order to stream distant sound sources to the HIs. Nowadays, microphones are the core sensors employed by the HIs to capture the sound. The technology behind them is relatively mature, being able to eliminate and overcome limitations of the past (Killion et al. 2016). Current microphones

are robust and are available in extremely small packages, especially in the case of micro-electrical-mechanical systems (MEMS) microphones.

In the future, more information about the acoustic environment might be retrieved by the use of remote microphones that are placed in the same space as the listener, such as microphones in laptops, mobile phones, TV sets and/or smart electronic appliances. The already existing wireless connectivity technologies of HIs can be employed for this purpose. This, together with audio-sharing networks will initiate a new era in audio signal processing and recognition, potentially enabling smarter and even more effective signal-processing schemes for the hearing impaired—but not only for them. Apart from smarter signal-processing schemes as are continuously evolving using additional microphones, information from other types of sensors may be exploited. In the next sections, an overview of some potential sensor applications is given.

4.1 Streaming Sound from Remote Sensors

Wireless systems are assistive listening devices that improve speech understanding achieved by the use of HIs. They can be used in various listening conditions, such as when the distance between the speaker and the listener is large, or when the hearing aids cannot provide sufficient speech-intelligibility enhancement—for instance, due to complex listening environments or severe hearing loss. The objective of wireless systems is to transmit a speech signal as clean as possible, without the undesired effects of noise and reverberation. The four major technologies driving assistive listening devices are infrared, induction, frequency of digital modulation systems, and Bluetooth (Staab 2013). In this section, two remote-microphone systems, known as FM systems, or digital modulation (DM) systems, are mainly considered.

A typical remote-microphone system consists of a small transmitter microphone that picks up the voice of a speaker, and sends the speech signal wirelessly to a RF receiver plugged or integrated into the HIs of a listener. The principle is shown on Fig. 5. The common use cases of such systems include classrooms, lecture halls, auditoria or restaurants (Staab 2013). The objective is to ensure a high-quality reproduction of the sound, whatever the distance between the speaker and the hearing-impaired listeners is. With a pure acoustic transmission, the sound intensity diminishes when the distance increases. The consequence is that both the SNR and DRR decrease—see Fig. 5. Remote microphones systems pick up the voice of the speaker close to their mouth, so that only the direct sound is recorded at a high SNR. Many studies have shown a strong intelligibility enhancement obtained with these systems when used with hearing aids (Hawkins 1984; Crandell and Smaldino 1999; Lewis et al. 2004; Thibodeau 2010; Schafer et al. 2013; Thibodeau 2014) or CIs (Wolfe et al. 2015; Vroegop et al. 2017).

Current remote-microphone systems provide a monophonic speech signal that is delivered to both ears. Although this is useful for speech intelligibility, the downside is that no spatial cues are reproduced. This may give rise to a feeling of isolation,

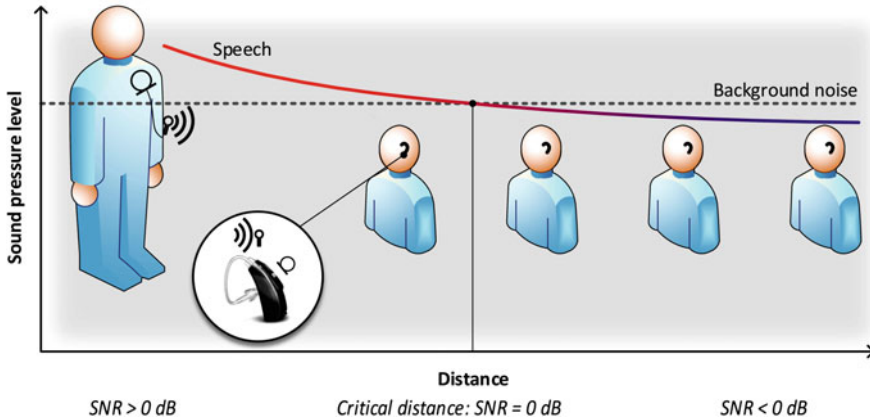


Fig. 5 Typical application of a remote-microphone system in a classroom

and is also prejudicial for speaker identification, for instance, in a multi-talker scenario with multiple remote microphones, or when the speaker is moving about in the room. The lack of spatial information can be partially solved in a reproduction mode where the clean transmitted speech is a mix of the monophonic clean speech and the degraded signal from the HI microphone(s)—however, at the cost of lower intelligibility (Thibodeau 2010).

The aforementioned issues may be solved with a different approach that consists in localizing the speaker in real time, as reported in Sect. 3.2, and then reproducing the transmitted speech after re-inserting the corresponding spatial cues (Courtois et al. 2015a). Another approach could be to generate a suitable ILD between the left and right streamed speech (Edwards 2016) and create a binaural signal using HRTFs stored in the devices (Aldaz et al. 2015; Courtois et al. 2016; Pontoppidan 2017). The HRTFs can be estimated online by examining the correlation between the remote microphone and the HI-microphone signals (Gran and Udesen 2017). Further, the spatialization can be combined with artificial reverberation according to the specific acoustical environment (Recker and Durant 2017)—thus generating kind of an *auditory virtual environment* for the HI user. Courtois et al. (2018) report a clinical study that assessed such a processing with 40 listeners. The results are in support of the integration of such solutions in future remote-microphone systems.

4.2 Gyroscopes, Accelerometers and Eye Trackers

Using sensors like gyroscopes and accelerometers, the head movements of the HI user can be tracked. Moreover, eye trackers based on electrooculography (EOG) sensors or on electrodes placed in the ear canal (Favre-Félix et al. 2017) can be used to detect where the listener is looking at. This information is useful for optimal tuning

of the sound-cleaning algorithms, since the target and noise signals may be localized more accurately (Boyd et al. 2013; Favre-Félix et al. 2018). Another aspect that could be improved by the use of such sensors is the spatial perception of streamed sound sources, which usually lack position information—compare Sect. 4.1. Furthermore, accelerometers and eye trackers can help recognizing the sound sources that HI users are currently attending to (Hadar et al. 1985; Tessendorf et al. 2011a) or to detect the current activities of the user, such as reading, walking, running. In addition, this information is useful for improving the parametrization of the algorithms and to control the switching between the various HI programs. By the use of accelerometers, further information related to the medical condition of the user may also get tracked—such as fall detection (Bagala et al. 2012; Thammasat and Chaicharn 2012; Wu et al. 2015) and/or gait abnormalities (Yang et al. 2009), to the end of potentially activating an emergency signal or triggering a real-time call for help.

4.3 Location-Detection (GPS) Sensors

Location-aware sensors, such as GPS sensors, may be also employed to capture user locations in order to enhance the automatic hearing-program selection (Tessendorf et al. 2011b). For example, the automatic-switching algorithm can consider whether the users are listening to an open air concert or they are just walking in the park. Personalized settings for specific environments may be remembered in this context. The GPS signals can also be used for keeping track of the location of the hearing instruments and prevent the loss of a device.

4.4 Electroencephalogram (EEG)

Capturing the users' auditory selective attention helps to recognize their current attended target. A practical problem in this context is that HIs do not know which source the user wants to attend to at a specific time. There is evidence that brain activity and the corresponding evoked electrical responses change as a function of which sound source the person is currently attending (Shinn-Cunningham and Best 2008; Mandic et al. 2010; Mesgarani and Chang 2012). Fuglsang et al. (2017) suggested that the tracking of an attended speech signal is performed in the cortex in a way that is invariant to acoustic distortions encountered in real-life sound environments. Thus, in theory, it should be possible to control the steering of the algorithms based on the tracked brain activity. This could be potentially achieved using electroencephalogram sensors placed in the ear canal (earEEG). EEG methods may also be used for fatigue monitoring, hearing-threshold detection or to monitor the physical condition of the user, for instance, respiratory and cardiovascular activity or sleeping (Looney et al. 2014, 2016). There are currently various ongoing research activities to explore the

feasibility of such approaches and of cognitively controlled HIs (Rodriguez-Villegas et al. 2010; Looney et al. 2014; Bleichner and Debener 2017; Lorenz et al. 2017).

4.5 Further Sensors

Other sensors that could be potentially integrated in HIs could be ear-worn ballistocardiogram sensors for monitoring the cardiac system’s condition or stress levels, for example, by measuring heart rates and cardiac contractility (He et al. 2012). Also, body-temperature, skin-conductance or blood-pressure sensors can provide information on the health status of the user.

4.6 Prognostication

It is challenging to predict the future with respect to the sensor availability within the applications of HIs. A list of potential sensor types and their application is given in Table 5. The information from additional sensors can enhance the processing performance of the instruments and could, potentially, allow for further functionality

Table 5 Sensor types and potential future applications

Sensor types	Potential applications
Gyroscope/accelerometer	HI user hearing wish detection
	Noise source localization
	Fall detection/gait abnormalities
	Automatic program selection
Eye-tracker	Target sound source detection
GPS sensor	Automatic program selection
	Lost device detection
	Retrieval of personalized settings
Electroencephalogram (EEG)	Intention
	Auditory attention
	Fatigue monitoring
	Hearing threshold detection
	Health condition monitoring
Heart-rate sensor	Stress
	Health condition monitoring
Temperature sensor	Health condition monitoring

related to the physical or the medical condition of the user, thus complementing the devices with health-monitoring applications (Tessendorf et al. 2013).

5 Conclusion

This chapter provided an overview of the current trends and challenges in the field of HI applications. It is evident that HIs of today have become intelligent systems, which offer processing strategies that are tailored to the individual patient and to specific environments. For this purpose, current HI algorithmic approaches have to take cognitive aspects into account, such as the listener intention, auditory attention, and the cognitive load. By appropriate parametrization and combined activation of existing hearing systems it is tried to deliver a sound that does not contradict the aforementioned cognitive functions. There are continuous research efforts going on to amend the knowledge base necessary for achieving optimal system settings, tailored to the current listening tasks of the HI users.

In fact, over the last years, there has been a growing interest within the research community in cognitive aspects related to hearing impairment and auditory attention. For instance, there are many efforts in trying to better understand the cognitive mechanisms of auditory perception and speech understanding. Moreover, technological advances related to wearables and the miniaturization of sensors will allow future HIs to be connected to the Internet and to other electronic devices. This progress will include the use of additional sensors, such as support microphones, accelerometers, head-trackers, pulse-rate meters, and EEG. Deeper knowledge in cognitive mechanisms and the additional information obtained from additional sensors will lead to improved parametrization of the signal-processing algorithms. Also, novel biologically inspired algorithms will, hopefully, be exploited for a further improvement of the daily-life situations of hearing-impaired people.

Acknowledgements The authors thank T. May and B. Kowalewski for their helpful comments. They further thank two anonymous reviewers for very constructive advice.

References

- Aldaz, G., M.B. Pedersen, M. Bergmann, S.O. Petersen, R.K. Pedersen, and P. Sommer. 2015. External microphone array and hearing aid using it. In *European Patents*, EP2840807 A1 (European Patent Office).
- Anemüller, J., and H. Kayser. 2017. Acoustic source localization by combination of supervised direction-of-arrival estimation with disjoint component analysis. In *International Conference on Latent Variable Analysis and Signal Separation*, ed. P. Tichavský, M. Babaie-Zadeh, O.J.J. Michel, and N. Thirion-Moreau, 99–108. Berlin: Springer.
- Appleton, J., and G. König. 2014. Improvement in speech intelligibility and subjective benefit with binaural beamformer technology. *Hearing Review*, Tech Topic Nov. 14.

- Bagala, F., C. Becker, A. Cappello, L. Chiari, K. Aminian, J.M. Hausdorff, W. Zijlstra, and J. Klenk. 2012. Evaluation of accelerometer-based fall detection algorithms on real-world falls. *PLOS ONE* 7 (5): e37062.
- Bishop, C. 2006. *Information Science and Statistics Pattern Recognition and Machine Learning*, 1st ed. Berlin: Springer.
- Blauert, J., and G.J. Brown. 2020. Reflective and reflexive auditory feedback. In *The Technology of Binaural Understanding*, eds. J. Blauert, and J. Braasch, 3–31. Cham, Switzerland: Springer and ASA press.
- Bleichner, M.G., and S. Debener. 2017. Concealed, unobtrusive ear-centered EEG acquisition: cEEGrids for transparent EEG. *Frontiers in Human Neuroscience* 11: 163.
- Bodden, M. 1993. Modeling human sound-source localization and the cocktail-party-effect. *Acta Acustica united with Acustica* 1: 43–55.
- Boyd, A.W., W.M. Whitmer, W.O. Brimijoin, and M.A. Akeroyd. 2013. Improved estimation of direction of arrival of sound sources for hearing aids using gyroscopic information. *Meetings on Acoustics (ICA)* 19: 030046.
- Braun, S., M. Torcoli, D. Marquardt, E.A. Habets, and S. Doclo. 2014. Multichannel dereverberation for hearing aids with interaural coherence preservation. In: *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, 124–128.
- Braun, S., W. Zhou, and E.A.P. Habets. 2015. Narrowband direction-of-arrival estimation for binaural hearing aids using relative transfer functions. In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, 1–5.
- Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge: MIT Press.
- Britannica. 2017. Deafness. In *Encyclopedia Britannica* (Encyclopedia Britannica Inc.).
- Büchler, M., S. Allegro, S. Launer, and N. Dillier. 2005. Sound classification in hearing aids inspired by auditory scene analysis. *EURASIP Journal on Advances in Signal Processing* 2005 (18): 387845.
- Capon, J. 1969. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE* 57 (8): 1408–1418.
- Cornelis, B., M. Moonen, and J. Wouters. 2014. Reduced-bandwidth multi-channel Wiener filter based binaural noise reduction and localization cue preservation in binaural hearing aids. *Signal Processing* 99: 1–16.
- Courtois, G. 2016. Spatial hearing rendering in wireless microphone systems for binaural hearing aids. PhD thesis, Ecole Polytechnique Fédérale de Lausanne.
- Courtois, G., H. Lissek, P. Estoppey, Y. Oesch, and X. Gigandet. 2018. Effects of binaural apatialization in wireless microphone systems for hearing aids on normal-hearing and hearing-impaired listeners. *Trends in Hearing* 22: 1–17.
- Courtois, G., P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande. 2014. Implementation of a binaural localization algorithm in hearing aids: Specifications and achievable solutions. In *Audio Engineering Society Convention 136*, Berlin, Germany.
- Courtois, G., P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande. 2015a. Binaural hearing aids with wireless microphone systems including speaker localization and spatialization. In *Audio Engineering Society Convention 138*, Warsaw, Poland.
- Courtois, G., P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande. 2015b. Development and assessment of a localization algorithm implemented in binaural hearing aids. In *23rd European Signal Processing Conference (EUSIPCO)*, Nice, France.
- Courtois, G., P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande. 2016. Hearing assistance systems. In *WHO Publications*, WO2016116160 A1 (World Health Organization (WHO)).
- Crandell, C.J., and J.J. Smaldino. 1999. Improving classroom acoustics: Utilizing hearing-assistive technology and communication strategies in the educational setting. *Volta Review* 101 (5): 47–62.
- Desloge, J.G., W.M. Rabinowitz, and P.M. Zurek. 1997. Microphone-array hearing aids with binaural output. I. Fixed-processing systems. *IEEE Transactions on Speech and Audio Processing* 5 (6): 529–542.

- Dillon, H. 2012. *Hearing Aids*. New York: Thieme.
- Doclo, S., S. Gannot, M. Moonen, and A. Spriet. 2010. Acoustic beamforming for hearing aid applications. In *Handbook on Array Processing and Sensor Networks*, 269–302. Hoboken: Wiley.
- Doclo, S., T.J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen. 2006. Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions. In *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Xian, China.
- Doclo, S., and M. Moonen. 2002. GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing* 50 (9): 2230–2244.
- Edwards, B. 2007. The future of hearing aid technology. *Trends in Amplification* 11 (1): 31–46.
- Edwards, B. 2016. Method and apparatus for a binaural hearing assistance system using monaural audio signals. In *US Patents*, US9510111 B2 (US Patent Office).
- Elko, G.W., and J. Meyer. 2008. Microphone arrays. In *Springer Handbook of Speech Processing, Springer Handbooks*. Berlin: Springer.
- Ellis, D., T. Virtanen, M.D. Plumbley, and B. Raj. 2018. Future perspective. In *Computational Analysis of Sound Scenes and Events*, 401–415. Berlin: Springer International Publishing.
- Faller, C., and J. Merimaa. 2004. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *Journal of the Acoustical Society of America* 116 (5): 3075–3089.
- Farmani, M., M.S. Pedersen, Z.H. Tan, and J. Jensen. 2015. Maximum likelihood approach to “informed” sound source localization for hearing aid applications. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Queensland, Australia, 16–20.
- Favre-Félix, A., C. Graversen, T. Dau, and T. Lunner. 2017. Real-time estimation of eye gaze by in-ear electrodes. In *International Engineering in Medicine and Biology Conference (EMBC)*, Jeju Island, Korea, 4086–4089.
- Favre-Félix, A., R. Hietkamp, C. Graversen, T. Dau, and T. Lunner. 2018. Steering of audio input in hearing aids by eye gaze through electrooculogram. In *ARO Midwinter Meeting*, San Diego, California.
- Froehlich, M., K. Freels, and T.A. Powers. 2015. Speech recognition benefit obtained from binaural beamforming hearing aids: Comparison to omnidirectional and individuals with normal hearing. *Audiology Online* 14338: 1–8.
- Frost, O.L. 1972. An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE* 60 (8): 926–935.
- Fuglsang, S.A., T. Dau, and J. Hjortkjær. 2017. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *Neuroimage* 156: 435–444.
- Geetha, C., K. Tanniru, and R.R. Rajan. 2017. Efficacy of directional microphones in hearing aids equipped with wireless synchronization technology. *The Journal of International Advanced Otology* 13 (1): 113–117.
- Gran, K.-F.J., and J. Udesen. 2017. Method of superimposing spatial auditory cues on externally picked-up microphone signals. In *US Patents*, US9699574 B2 (US Patent Office).
- Griffiths, L., and C.W. Jim. 1982. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation* 30 (1): 27–34.
- Hadad, E., S. Doclo, and S. Gannot. 2016. The binaural LCMV beamformer and its performance analysis. *IEEE Transactions on Audio, Speech and Language Processing* 24 (3): 543–558.
- Hadad, E., S. Gannot, and S. Doclo. 2012. Binaural linearly constrained minimum variance beamformer for hearing aid applications. In *13th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany, 1–4.
- Hadar, U., T.J. Steiner, and F. Clifford Rose. 1985. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior* 9 (4): 214–228.
- Hamacher, V., J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass. 2005. Signal processing in high-end hearing aids: State of the art, challenges, and future trends. *EURASIP Journal on Advances in Signal Processing* 2005: 2915–2929.

- Hamacher V., E. Fischer, U. Kornagel, and H. Puder. 2006. Applications of adaptive signal processing methods in high-end hearing aids. In *Topics Acoustic Echo and Noise Control*, 599–636. Berlin: Springer Science & Business Media.
- Hassager, H.G., T. May, A. Wiinberg, and T. Dau. 2017a. Preserving spatial perception in rooms using direct-sound driven dynamic range compression. *Journal of the Acoustical Society of America* 141 (6): 4556–4566.
- Hassager, H.G., A. Wiinberg, and T. Dau. 2017b. Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment. *Journal of the Acoustical Society of America* 141 (4): 2556–2568.
- Hawkins, D.B. 1984. Comparisons of speech recognition in noise by mildly-to-moderately hearing-impaired children using hearing aids and FM systems. *The Journal of Speech and Hearing Disorders* 49 (4): 409–418.
- He, D.D., E.S. Winokur, and C.G. Sodini. 2012. An ear-worn continuous ballistocardiogram (BCG) sensor for cardiovascular monitoring. In *International Engineering in Medicine and Biology Conference (EMBC)*, San Diego, USA, vol. 2012, 5030–5033.
- Hiruma, N., H. Nakashima, and Y.-I. Fujisaka. 2016. Low delay wind noise cancellation for binaural hearing aids. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, Hamburg, Germany, vol. 253, 4844–4854.
- Holube, I., H. Pudder, and T. Velde. 2014. DSP hearing instruments. In *Sandlin's Textbook of Hearing Aid Implication - Technical and Clinical Considerations*, 221–293. San Diego, CA: Plural Publishing.
- Ibrahim, I., V. Parsa, E. Macpherson, and M. Cheesman. 2013. Evaluation of speech intelligibility and sound localization abilities with hearing aids using binaural wireless technology. *Audiology Research* 3 (1): 1–21.
- Kates, J.M. 1995. Classification of background noises for hearing aid applications. *Journal of the Acoustical Society of America* 97 (1): 461–470.
- Kates, J.M. 2008. *Digital Hearing Aids*. San Diego: Plural Publishing.
- Keidser, G., A. O'Brien, J.-U. Hain, M. McLelland, and I. Yeend. 2009. The effect of frequency-dependent microphone directionality on horizontal localization performance in hearing-aid users. *International Journal of Audiology* 48 (11): 789–803.
- Keidser, G., K. Rohrseitz, H. Dillon, V. Hamacher, L. Carter, U. Rass, and E. Convery. 2006. The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers. *International Journal of Audiology* 45 (10): 563–579.
- Killion, M., A. van Halteren, S. Stenfelt, and D. Warren. 2016. Hearing aid transducers. In *Hearing Aids*, 59–92. Berlin: Springer.
- Killion, M.C. 1979. AGC circuit particularly for a hearing aid. In *US Patents*, US4170720 A (US Patent Office).
- Klasen, T.J., M. Moonen, T. Van den Bogaert, and J. Wouters. 2005. Preservation of interaural time delay for binaural hearing aids through multi-channel Wiener filtering based noise reduction. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Queensland, Australia, vol. 3, iii–29.
- Klasen, T.J., T. Van den Bogaert, M. Moonen, and J. Wouters. 2007. Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues. *IEEE Transactions on Signal Processing* 55 (4): 1579–1585.
- Kollmeier, B., and R. Koch. 1994. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *Journal of the Acoustical Society of America* 95 (3): 1593–1602.
- Kollmeier, B., J. Peissig, and V. Hohmann. 1993. Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain. *Scandinavian Audiology Supplementum* 38: 28–38.

- Korhonen, P., C. Lau, F. Kuk, D. Keenan, and J. Schumacher. 2015. Effects of coordinated compression and pinna compensation features on horizontal localization performance in hearing aid users. *Journal of the American Academy of Audiology* 26 (1): 80–92.
- Latzel, M. 2013. Concepts for binaural processing in hearing aids. *Hearing Review, Hearing Instruments*, March 2013. <http://www.hearingreview.com/2013/03/concepts-for-binaural-processing-in-hearing-aids/>. (last accessed December 20, 2019).
- Laurer, S., J.A. Zakis, and B.C. Moore. 2016. Hearing aid signal processing. In *Hearing Aids*, 93–130. Berlin: Springer.
- Lewis, M.S., C.C. Crandell, M. Valente, and J.E. Horn. 2004. Speech perception in noise: Directional microphones versus frequency modulation (FM) systems. *Journal of the American Academy of Audiology* 15 (6): 426–439.
- Liao, W.-C., M. Hong, I. Merks, T. Zhang, and Z.-Q. Luo. 2015a. Incorporating spatial information in binaural beamforming for noise suppression in hearing aids. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Queensland, Australia, 5733–5737.
- Liao, W.-C., Z.-Q. Luo, I. Merks, and T. Zhang. 2015b. An effective low complexity binaural beamforming algorithm for hearing aids. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, 1–5.
- Looney, D., V. Goverdovsky, I. Rosenzweig, M.J. Morrell, and D.P. Mandic. 2016. Wearable in-ear encephalography sensor for monitoring sleep. Preliminary observations from nap studies. *Annals of the American Thoracic Society* 13 (12): 2229–2233.
- Looney, D., P. Kidmose, and D.P. Mandic. 2014. Ear-EEG: User-centered and wearable bci. In *Brain-Computer Interface Research: A State-of-the-Art Summary -2*, 41–50. Berlin: Springer.
- Lorenz, F., W. Malte, G. Carina, B. Alex, L. Thomas, and O. Jonas. 2017. Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. *Journal of Neural Engineering* 14 (3): 036020.
- Ma, N., T. May, and G.J. Brown. 2017. Exploiting Deep Neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE Transactions on Acoustics Speech and Signal Processing* 25 (12): 2444–2453.
- Mandic, P.K., M.L. Rank, M. Ungstrup, D. Looney, C. Park, and P., D. 2010. A yabus-style experiment to determine auditory attention. In *International Engineering in Medicine and Biology Conference (EMBC)*, Buenos Aires, Argentina, 4650–4653.
- Marquardt, D., E. Hadad, S. Gannot, and S. Doclo. 2014. Optimal binaural LCMV beamformers for combined noise reduction and binaural cue preservation. In *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Juan-les-Pins, France, 288–292.
- Marquardt, D., V. Hohmann, and S. Doclo. 2013. Coherence preservation in multi-channel Wiener filtering based noise reduction for binaural hearing aids. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 8648–8652.
- Marquardt, D., V. Hohmann, and S. Doclo. 2015. Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids. *IEEE Transactions on Audio, Speech and Language Processing* 23 (12): 2162–2176.
- May, T. 2018. Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues. *IEEE Transactions on Audio, Speech and Language Processing* 26 (2): 406–414.
- May, T., S. van de Par, and A. Kohlrausch. 2013. Binaural localization and detection of speakers in complex acoustic scenes. In *The Technology of Binaural Listening*, ed. Jens Blauert, 397–425. Springer and ASA Press.
- May, Tobias , Borys Kowalewski, and Torsten Dau. 2020. Scene-aware dynamic range compression in hearing aids. In: *The Technology of Binaural Understanding*, eds. J. Blauert and J. Braasch, 763–799. Cham, Switzerland: Springer and ASA press.
- Merks, I., G. Enzner, and T. Zhang. 2013. Sound source localization with binaural hearing aids using adaptive blind channel identification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 438–442.

- Mesgarani, N., and E.F. Chang. 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485 (7397): 233–236.
- Mollick, E. 2006. Establishing Moore's law. *IEEE Annals of the History of Computing* 28 (3): 62–75.
- Moore, B.C. 2007a. Binaural sharing of audio signals: Prospective benefits and limitations. *The Hearing Journal* 60 (11): 46–48.
- Moore, B.C. 2007b. Hearing aids. In *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*. Hoboken: Wiley.
- Moradi, S., B. Lidestam, H. Danielsson, E.H.N. Ng, and J. Rönnerberg. 2017. Visual cues contribute differentially to audiovisual perception of consonants and vowels in improving recognition and reducing cognitive demands in listeners with hearing impairment using hearing aids. *Journal of Speech, Language, and Hearing Research* 60 (9): 2687–2703.
- Neher, T., K.C. Wagener, and R.-L. Fischer. 2016. Directional processing and noise reduction in hearing aids: Individual and situational influences on preferred setting. *Journal of the Acoustical Society of America* 27 (8): 628–646.
- Neher, T., K.C. Wagener, and M. Latzel. 2017. Speech reception with different bilateral directional processing schemes: Influence of binaural hearing, audiometric asymmetry, and acoustic scenario. *Hearing Research* 353: 36–48.
- Nishimura, R., Y. Suzuki, and F. Asano. 2002. A new adaptive binaural microphone array system using a weighted least squares algorithm. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, Florida, USA, vol. 2, II–1925.
- Nordqvist, P., and A. Leijon. 2004. An efficient robust sound classification algorithm for hearing aids. *Journal of the Acoustical Society of America* 115 (6): 3033–3041.
- Picou, E.M., E. Aspell, and T.A. Ricketts. 2014. Potential benefits and limitations of three types of directional processing in hearing aids. *Ear and Hearing* 35 (3): 339–352.
- Pontoppidan, N.H. 2017. Binaural hearing assistance system comprising a database of head related transfer functions. In *US Patents*, US9565502 B2 (US Patent Office).
- Popelka, G.R., and B.C.J. Moore. 2016. Future directions for hearing aid development. In *Hearing Aids*, 323–333. Berlin: Springer.
- Recker, K.L., and E.A. Durant. 2017. Method and apparatus for localization of streaming sources in hearing assistance system. In *US Patents*, US9584933 B2 (US Patent Office).
- Ricketts, T.A., E.M. Picou, and J. Galster. 2017. Directional microphone hearing aids in school environments: Working toward optimization. *Journal of Speech, Language, and Hearing Research* 60 (1): 263–275.
- Rodriguez-Villegas, A.J.C., D.C. Yates, S.J.M. Smith, and J.S. Duncan. 2010. Wearable electroencephalography. *IEEE Engineering in Medicine and Biology Magazine* 29 (3): 44–56.
- Schafer, E.C., L. Mathews, S. Mehta, M. Hill, A. Munoz, R. Bishop, and M. Moloney. 2013. Personal FM systems for children with autism spectrum disorders (ASD) and/or attention-deficit hyperactivity disorder (ADHD): An initial investigation. *Journal of Communication Disorders* 46 (1): 30–52.
- Schwartz, A.H., and B.G. Shinn-Cunningham. 2013. Effects of dynamic range compression on spatial selective auditory attention in normal-hearing listeners. *Journal of the Acoustical Society of America* 133 (4): 2329–2339.
- Schwartz, B., S. Gannot, and E.A. Habets. 2015. An online dereverberation algorithm for hearing aids with binaural cues preservation. In *Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, 1–5.
- Shinn-Cunningham, B.G., and V. Best. 2008. Selective attention in normal and impaired hearing. *Trends in Amplification* 12 (4): 283–299.
- Smith, P., A. Davis, J. Day, S. Unwin, G. Day, and J. Chalupper. 2008. Real-world preferences for linked bilateral processing. *Hearing Journal* 61 (7): 33–34.
- Sockalingam, R., M. Holmberg, K. Eneroth, and M. Shulte. 2009. Binaural hearing aid communication shown to improve sound quality and localization. *Hearing Journal* 62 (10): 46–47.

- Soede, W., A.J. Berkhout, and F.A. Bilsen. 1993. Development of a directional hearing instrument based on array technology. *Journal of the Acoustical Society of America* 94 (2): 785–798.
- Souza, P. 2016. Speech perception and hearing aids. In *Hearing Aids*, 151–180. Berlin: Springer.
- Srinivasan, S., A. Pandharipande, and K. Janse. 2008. Beamforming under quantization errors in wireless binaural hearing aids. *Journal on Audio, Speech, and Music Processing* 2008 (1).
- Staab, W. 2013. Wireless systems for hearing aids. *Hearing Health & Technology Matters*. <http://hearinghealthmatters.org/waynesworld/2013/2566/>. (last accessed Decemebr 18, 2019).
- Stadler, R.W., and W.M. Rabinowitz. 1993. On the potential of fixed arrays for hearing aids. *Journal of the Acoustical Society of America* 94 (3): 1332–1342.
- Szurley, J., A. Bertrand, B. Van Dijk, and M. Moonen. 2016. Binaural noise cue preservation in a binaural noise reduction system with a remote microphone signal. *IEEE Transactions on Audio, Speech and Language Processing* 24 (5): 952–966.
- Tessendorf, B., A. Bulling, D. Roggen, T. Stiefmeier, M. Feilner, P. Derleth, and G. Tröster. 2011a. Recognition of hearing needs from body and eye movements to improve hearing instruments. In *International Conference on Pervasive Computing*, San Francisco, USA, 314–331.
- Tessendorf, B., A. Kettner, D. Roggen, T. Stiefmeier, G. Tröster, P. Derleth, and M. Feilner. 2011b. Identification of relevant multimodal cues to enhance context-aware hearing instruments. In *International Conference on Body Area Networks*, Beijing, China, 15–18.
- Tessendorf, B., M. Debevc, P. Derleth, M. Feilner, F. Gravenhorst, D. Roggen, T. Stiefmeier, and G. Tröster. 2013. Design of a multimodal hearing system. *Computer Science and Information Systems* 10 (1).
- Thammasat, E., and J. Chaicharn. 2012. A simply fall-detection algorithm using accelerometers on a smartphone. In *Biomedical Engineering International Conference*, Penang, Malaysia, 1/4.
- Thibodeau, L. 2010. Benefits of adaptive FM systems on speech recognition in noise for listeners who use hearing aids. *American Journal of Audiology* 19 (1): 36–45.
- Thibodeau, L. 2014. Comparison of speech recognition with adaptive digital and FM remote microphone hearing assistance technology by listeners who use hearing aids. *American Journal of Audiology* 23 (2): 201–210.
- Thiemann, J., M. Müller, D. Marquardt, S. Doclo, and S. van de Par. 2016. Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene. *EURASIP Journal on Advances in Signal Processing* 2016 (1): 12.
- Timmer, B. 2013. Is it sync or stream? The differences between wireless hearing aid features. *The Hearing Review* 20 (6): 20–22.
- Tsilfidis, A., A. Westermann, J.M. Buchholz, E. Georganti, and J. Mourjopoulos. 2013. *Binaural Dereverberation*, 359–396. Berlin: Springer.
- Van den Bogaert, T., E. Cayette, and J. Wouters. 2011. Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna. *International Journal of Audiology* 50 (3): 164–176.
- Varghese, L.A., E.J. Ozmeral, V. Best, and B.G. Shinn-Cunningham. 2012. How visual cues for when to listen aid selective auditory attention. *Journal of the Association for Research in Otolaryngology* 13 (3): 359–368.
- Vroegop, J.L., J.G. Dingemanse, N.C. Homans, and A. Goedegebure. 2017. Evaluation of a wireless remote microphone in bimodal cochlear implant recipients. *International Journal of Audiology* 56 (9): 643–649.
- Wang, D., and G.J. Brown. 2007. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. New York: IEEE Press/Wiley-Interscience.
- Welker, D.P., J.E. Greenberg, J.G. Desloge, and P.M. Zurek. 1997. Microphone-array hearing aids with binaural output. II. A two-microphone adaptive system. *IEEE Transactions on Speech and Audio Processing* 5 (6): 543–551.
- Westermann, A., J.M. Buchholz, and T. Dau. 2013. Binaural dereverberation based on interaural coherence histograms. *Journal of the Acoustical Society of America* 133 (5): 2767–2777.
- WHO. 2017. Deafness and hearing loss. <http://www.who.int/mediacentre/factsheets/fs300/en/>. (last accessed December 13, 2019).

- Widrow, B., J.R. Glover, J.M. McCool, J. Kaunitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, J.E. Dong, and R.C. Goodlin. 1975. Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE* 63 (12): 1692–1716.
- Widrow, B., and F.-L. Luo. 2003. Microphone arrays for hearing aids: An overview. *Speech Communication* 39 (1): 139–146.
- Wiggins, I.M., and B.U. Seeber. 2012. Effects of dynamic-range compression on the spatial attributes of sounds in normal-hearing listeners. *Ear and Hearing* 33 (3): 399–410.
- Wilson, K., and T. Darrell. 2005. Improving audio source localization by learning the precedence effect. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Pennsylvania, USA, vol. 4, iv/1125–iv/1128.
- Wittkop, T., V. Hohmann, and B. Kollmeier. 1996. Noise reduction strategies in digital binaural hearing aids. In *International Symposium on Psychoacoustics, Speech and Hearing Aids*, 245–251.
- Wolfe, J., M. Morais, E. Schafer, S. Agrawal, and D. Koch. 2015. Evaluation of speech recognition of cochlear implant recipients using adaptive, digital remote microphone technology and a speech enhancement sound processing algorithm. *Journal of the American Academy of Audiology* 26 (5): 502–508.
- Wu, F., H. Zhao, Y. Zhao, and H. Zhong. 2015. Development of a wearable-sensor-based fall detection system. *International Journal of Telemedicine and Applications* 2015: 2.
- Wu, Y.H., and R.A. Bentler. 2010. Impact of visual cues on directional benefit and preference: Part i. Laboratory tests. *Ear and Hearing* 31 (1): 22–34.
- Yang, C.-Y., W.-S. Chou, K.-C. Chang, C.-W. Liu, T.-S. Chi, and S.-L. Jou. 2013. Spatial-cue-based multi-band binaural noise reduction for hearing aids. In *Workshop on Signal Processing Systems (SiPS)*, Taipei City, Taiwan, 278–283.
- Yang, L.A., O. Aziz, B. Lo, and Z., G. 2009. Detecting walking gait impairment with an ear-worn sensor. In *International Workshop on Wearable and Implantable Body Sensor Networks*, Berkeley, California, USA, 175–180.
- Yee, D., H. Kamkar-Parsi, R. Martin, and H. Puder. 2017. A noise reduction post-filter for binaurally-linked single-microphone hearing aids utilizing a nearby external microphone. *IEEE Transactions on Acoustics Speech and Signal Processing* 26 (1): 5–18.
- Yousefian, N., P.C. Loizou, and J.H. Hansen. 2014. A coherence-based noise reduction algorithm for binaural hearing aids. *Speech Communication* 58: 101–110.
- Zohourian, M., and R. Martin. 2016. Binaural speaker localization and separation based on a joint ILD/ITD model and head movement tracking. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 430–434.

Scene-Aware Dynamic-Range Compression in Hearing Aids



Tobias May, Borys Kowalewski and Torsten Dau

Abstract Wide dynamic-range compression (WDRC) is one of the essential building blocks in hearing aids and aims at improving audibility while maintaining acceptable loudness at high sound pressure levels for hearing-impaired (HI) listeners. While fast-acting compression with a short release time allows amplifying low-intensity speech sounds on short time scales corresponding to syllables or phonemes, such processing also typically amplifies noise components in speech gaps. The latter reduces the output signal-to-noise ratio (SNR) and disrupts the acoustic properties of the background noise. Moreover, the use of fast-acting compression distorts auditory cues involved in the spatial perception of sounds in rooms by amplifying low-level reverberant energy portions of the sound relative to the direct sound. Some of these shortcomings can be avoided by choosing a longer release time, but such a slow-acting compression system fails to amplify soft speech components on short time scales and compromises on the ability to restore loudness perception. This chapter investigates the benefit of a new scene-aware dynamic-range compression strategy, which attempts to combine the advantages of both fast- and slow-acting compression. Specifically, the release time of the compressor is adaptively changed to provide fast- and slow-acting compression depending on whether the target was present or absent. The benefit of this scene-aware compression strategy was evaluated instrumentally in acoustic scenarios where speech and noise were present simultaneously. Moreover, a subjective listening test was conducted to assess the impact of scene-aware compression on reverberant speech signals by measuring the perceived location and spatial distribution of virtualized speech in normal-hearing (NH) listeners.

T. May (✉) · B. Kowalewski · T. Dau
Hearing Systems Section, Department of Health Technology,
Technical University of Denmark, 2800 Lyngby, Denmark
e-mail: tobmay@dtu.dk

© Springer Nature Switzerland AG 2020
J. Blauert and J. Braasch (eds.), *The Technology of Binaural Understanding*,
Modern Acoustics and Signal Processing,
https://doi.org/10.1007/978-3-030-00386-9_25

1 Introduction

A well-functioning cochlea acts like a frequency-selective, level-dependent amplifier. The passive mechanical tuning is enhanced due to the action of the outer hair cells (Robles and Ruggero 2001). As the stimulus intensity decreases, the outer hair cells provide increasing gain. This means that each region of the basilar membrane of the cochlea elicits a nonlinear compressive input/output function (Oxenham and Bacon 2004). Sensorineural hearing loss is in many cases associated with an impairment of this active cochlear mechanism due to, e.g., exposure to noise, the use of ototoxic drugs, aging, or a disease (Kramer 2008). As a consequence, hearing loss inevitably leads to an incomplete, distorted internal representation of the sound (Lopez-Poveda 2014). HI listeners often have difficulties with speech recognition, especially in noisy or reverberant conditions or when many talkers are involved. Even with hearing-aid amplification, they might not achieve the same performance as NH listeners (Souza 2016). Spatial (or directional) hearing abilities may also become impaired in listeners with a hearing loss, which may further contribute to their poor speech recognition performance in adverse conditions (e.g., Noble et al. 1995, 1997; Keidser et al. 2006). Age-related central processing deficits and cognitive impairments may also contribute to poor speech intelligibility performance. Nevertheless, impairments stemming from the auditory periphery are often considered a primary source of the difficulties experienced by HI listeners (Humes 2002).

Modern hearing aids provide a range of signal-processing algorithms, such as directional filtering (beamforming), noise reduction, and dynamic-range compression (see, e.g., Dillon 2008). The purpose of such hearing-aid algorithms is to improve speech intelligibility and listening comfort (Neher et al. 2014). However, since the primary consequence of a hearing loss is reduced audibility, amplification represents the most basic function of a hearing instrument (Souza et al. 2007). Early hearing aids were usually linear and provided constant gain independently of the input signal level. Since the individual sensitivity of hearing typically varies across frequency, the amount of gain would be adjusted individually for several frequency channels. However, the elevation of the threshold in quiet does not correlate with an equal increase of the uncomfortable level (UCL; Dillon and Storey 1998). As a result, many HI listeners' experience loudness recruitment and a reduced dynamic range of levels. To ensure a comfortable loudness perception of amplified speech, the gain profile as a function of frequency does not "mirror" the audiogram. In fact, most linear rationales prescribe a gain of 0.5 dB (or below) per 1 dB of hearing loss, reflecting the so-called *half-gain rule* (see Dillon 2008, for a review). Linear amplification would represent a compromise between audibility and loudness comfort (Villchur 1973; Edwards 2004). Therefore, nonlinear, level-dependent circuits have been proposed to avoid excessive loudness and compensate for the limited dynamic range (Villchur 1973; Barfod 1978; Kuk 1996; Souza 2002; Moore 2008). WDRC systems typically do that by providing a constant gain for input signals below a predefined level, known as the compression threshold (CT), and by reducing the gain for signal components above the CT. Such processing allows reducing the output dynamic range

on short time scales corresponding to syllables or phonemes, thus “squeezing” the speech information through the informational “bottleneck” created by the impaired auditory system (Kuk 1996).

It is widely accepted that compression can provide increased audibility of soft sound components while maintaining comfortable loudness (Kates 2010; Holube et al. 2016). This requires a level detection circuit that can follow fast level fluctuations in the input signal in several frequency channels, which justifies the use of multi-channel compression with short attack and release times. Such an approach, however, might also lead to distortions of the spectral and temporal envelopes of speech. Moreover, as demonstrated by Naylor and Johannesson (2009) and Hassager et al. (2017b), the fast gain fluctuations can cyclically amplify the background noise and/or reverberation, decreasing the SNR and the direct-to-reverberant energy ratio (DRR). Bilateral compression, operating independently in each ear, may also introduce unnatural fluctuations of the interaural level differences (ILDs) as shown by Musa-Shufani et al. (2006) and others. These types of distortions can severely disrupt the spatial perception of an acoustic scene. To avoid such compression-induced distortions, less aggressive parameters, such as longer time constants or lower compression ratios (CRs) could be used, but this effectively linearizes the system. Slow-acting compression can still adjust the long-term signal level to achieve desirable loudness, but it cannot follow fluctuations on time scales corresponding to phonemes or syllables.

Instead of using WDRC with a fixed set of parameters, it has been suggested that compression parameters should be adjusted dynamically depending on the current acoustic scenario (Gatehouse et al. 2006a; Souza et al. 2012a). Moreover, due to the dynamic nature of real-world signals, where the SNR can vary substantially across time, such adjustments should probably be made on relatively short time scales. Several adaptive compression strategies have been proposed that adjust the time constants according to changes in the input level or based on the current dynamic range (Killion et al. 1992; Lai et al. 2013). Moreover, based on the results from Hassager et al. (2017b), it seems beneficial to adjust the time constants depending on the estimated short-term SNR or the DRR in a manner similar to noise reduction and de-reverberation systems currently implemented in hearing aids.

This chapter provides an overview of state-of-the-art WDRC systems for hearing aid applications. First, the main building blocks of a conventional WDRC system are reviewed in Sect. 2. Then, the perceptual consequences of different WDRC system configurations are discussed in Sect. 3. Motivated by the limitations of a fixed set of WDRC parameters, the concept of scene-aware dynamic-range compression is presented in Sect. 4, where the characteristics of the compressor are adjusted in individual time-frequency (T-F) units depending on the presence of the target signal. Section 5 provides an overview of perceptually-relevant instrumental metrics, which can be used to evaluate the effects of WDRC processing on the speech signal and the background. In Sect. 6, two application scenarios of scene-aware compression are presented. First, a monaural SNR-aware WDRC system is described and evaluated

in a speech-in-noise scenario using a set of instrumental metrics. Second, results of a spatial-perception experiment are presented, in which a binaural DRR-aware compressor is compared to a conventional WDRC system using speech signals in a reverberant environment. Finally, the chapter is summarized in Sect. 7.

2 Dynamic-Range Compression

The major building blocks of a conventional WDRC system are shown in Fig. 1. The input signal is typically analyzed by a short-time discrete Fourier transform (STFT). Afterwards, a filterbank can be applied to group the individual discrete Fourier transform (DFT) bins into a predefined number of frequency channels, in which the short-term level is subsequently estimated. Based on this level estimation, a frequency-specific gain function is computed. Then, the channel-specific gain function is interpolated to individual DFT bins and applied to the STFT representation of the input signal. Finally, the processed output signal is reconstructed from the modified STFT representation by applying an inverse short-time discrete Fourier transform (ISTFT). All individual building blocks are described in the following subsections.

2.1 STFT Analysis

The input signal is divided into successive, overlapping segments. Each segment is then weighted by an analysis window and zero-padded to a length that typically corresponds to a next-higher power of two. Afterwards, the DFT is computed for each frame, producing the STFT representation of the input signal (Allen 1977). A short window duration between 8 and 16 ms with 50 or 75 % overlap is typically used to avoid a clearly audible delay, which would be disturbing (Stone and Moore 1999, 2002).

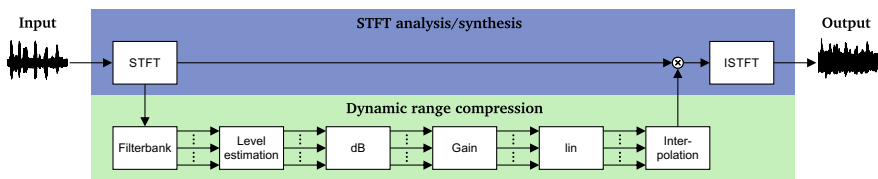


Fig. 1 Block diagram of a conventional WDRC system consisting of two processing layers: (1) analysis and synthesis stage based on the STFT and (2) dynamic range compression in individual frequency channels

2.2 Filterbank

Compression is more effective when operating in independent, narrow frequency channels (Henning and Bentler 2008; Naylor and Johannesson 2009). Moreover, frequency-dependent compression allows to better accommodate the variation of the hearing thresholds and the residual dynamic range across frequency in individual listeners and provides better speech audibility (Kuk 1996; Souza and Turner 1998; Souza 2002; Dillon 2008). Commercially-available hearing aids often offer up to 20 processing channels (Souza 2002; Cox et al. 2016). However, Woods et al. (2006) demonstrated that four channels are usually sufficient to obtain a good fit to the gain prescription targets. Woods et al. also suggested that considering the speech intelligibility index (SII), no more than five compression channels are necessary for a mild-to-moderate hearing loss and no more than nine channels for a more severe impairment. Studies of Yund and Buckles (1995b) and Alexander and Masterson (2014) suggested eight as an optimal number of channels for speech intelligibility, regardless of the SNR and the release time of the compressor. Further increasing the resolution of the filterbank seems to disrupt the spectral contrast of speech, which can be detrimental to speech intelligibility (De Gennaro et al. 1986; Bustamante and Braida 1987; Souza et al. 2005; Bor et al. 2008; Holube et al. 2016) and sound quality (van Buuren et al. 1999).

The STFT representation of the signal is used to create a number of frequency channels. This can be achieved by grouping the individual DFT bins and, optionally, by applying different weights to the bins within one group, in order to achieve a desired filter shape and bandwidth. In the example shown in Fig. 2 and in the experiments discussed in Sect. 6, the DFT bins are grouped together to create rectangular, octave-wide filters. An overview of alternative filterbanks with emphasis on low-delay implementations can be found in Kates (2005).

2.3 Level Estimation

The gain function of a WDRC system depends on the estimated level of the signal at the input to the system. To avoid rapid fluctuations of the gain function across time, typically some form of smoothing is applied. To achieve this, the short-term level which is estimated for each frame is usually smoothed across time by a first-order infinite impulse response (IIR) low-pass filter with different time constants associated with the attack and the release (Kates 1993). Alternatively, an instantaneous level estimator could be used and the smoothing with different attack and release time constants could be applied after the gain calculation stage (Giannoulis et al. 2012). The nominal time constants of the smoothing filter are usually referred to as the *RC time constants*. More often, however, effective time constants are reported, measured in response to a well-defined test signal of varying level. One such measurement procedure is described in the ANSI S3.22-1996 standard.

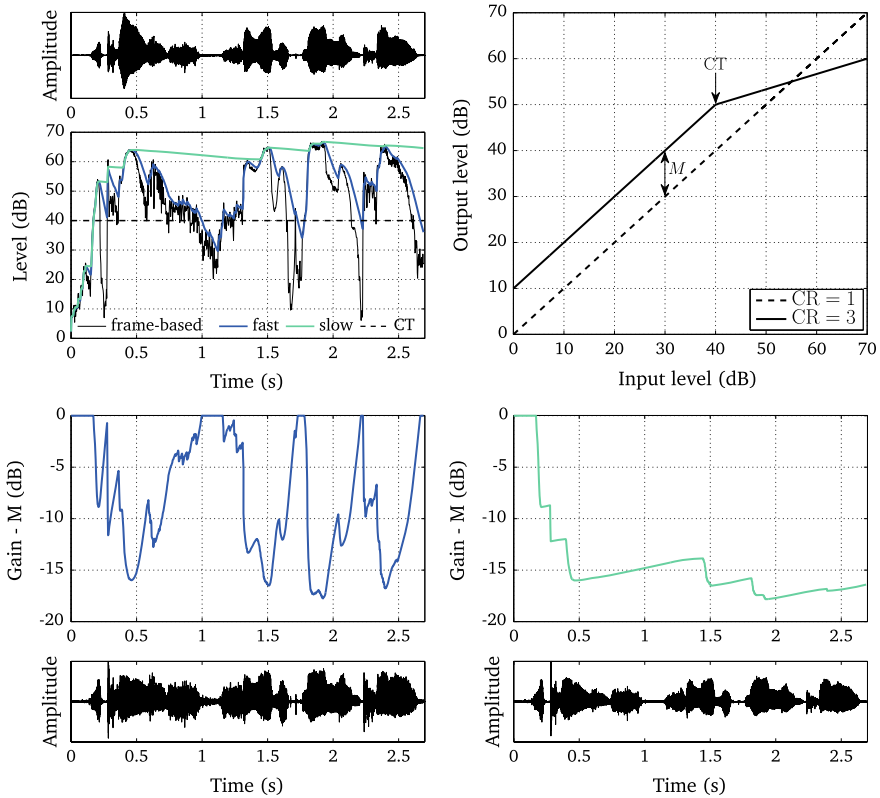


Fig. 2 Illustration of the main building blocks of a WDRC system: Input signal and level estimation for a frequency channel centered at 1 kHz using a 5 ms attack time and either a fast (40 ms) or a slow (2000 ms) release time (top left), static input/output function with a compression ratio of $CR = 3$ along with a make-up gain of $M = 10$ dB and a compression threshold of $CT = 40$ dB (top right), gain and respective output signal of a fast-acting compressor (bottom left), gain and respective output signal of a slow-acting compressor (bottom right)

To ensure that the compressor is able to respond to sudden level increases, a fast attack time between 1 and 5 ms is typically used while the choice of the release time determines whether the resulting system provides fast-acting (<200 ms) or slow-acting (>200 ms) WDRC compression (Souza 2002, time constants defined according to the ANSI standard). The influence of the release time on the level estimation is illustrated in Fig. 2, where a simplified WDRC system with channel-independent CTs and CRs was used. The perceptual effects of fast- and slow-acting WDRC are discussed in detail in Sect. 3.

2.4 Gain Calculation

In the gain calculation stage, a broken-stick gain function is typically used to map the estimated input level in decibels in individual frequency channels to a corresponding output level in decibels. As illustrated in the top right panel of Fig. 2, this mapping provides linear amplification for low-level inputs below the CT. This linear gain will be referred to as make-up gain M . Above the CT, this gain is reduced according to a constant CR. The CR and the make-up gain depend on the individual listener's audiogram and are usually determined by a nonlinear gain prescription rule, such as NAL-NL1/NL2 (Dillon 1999; Keidser et al. 2011), CAM2 (Moore and Sek 2013) or DSL [i/o] (Scollie et al. 2005). These prescription rules are typically optimized for speech intelligibility, using a metric such as the SII (ANSI S3.5 1997), while attempting to restore normal loudness perception and not exceeding the listener's uncomfortable level (UCL) at a given frequency. The CT for each frequency channel can be derived by averaging the short-term levels across time that were estimated in response to stationary speech-shaped noise at a predefined overall sound pressure level (SPL), typically around 40–50 dB SPL. However, even lower thresholds can be encountered in commercial hearing aids (Souza 2002).

2.5 Interpolation

The linear gain that was calculated for individual frequency channels has to be interpolated from channel center frequencies to DFT bin frequencies. Such a mapping can, for example, be accomplished by employing a piecewise cubic interpolation (Hassager et al. 2017a; May et al. 2018). The interpolated gain function can be applied subsequently to the STFT representation of the input signal. Ensuring that the frequency response of the compressor changes smoothly across frequency helps to avoid aliasing artifacts (Kates 2005).

2.6 STFT Synthesis

The output signal of the modified STFT representation can be obtained by applying the ISTFT. After calculating the inverse discrete Fourier transform (IDFT) for each frame, a synthesis window is usually applied to reduce discontinuities at the frame boundaries which can occur due to the application of the gain function. Afterward, the time domain signal can be reconstructed by the overlap-add (OLA) method (Allen 1977). A raised cosine window is typically used as a synthesis window (Grimm et al. 2006; Strelcyk et al. 2012).

2.7 Illustration of WDRC

The processing of a conventional WDRC system is illustrated in Fig. 2 for the first segment of the TIMIT sentence: “*Materials: ceramic modeling clay ...*”. The WDRC system used a window duration of 10 ms with 75 % overlap and operated in seven octave-wide frequency channels with center frequencies spaced between 125 Hz and 8 kHz. The input signal is shown in the top left panel of Fig. 2 along with estimated levels for a frequency channel centered at 1 kHz using the frame-based energy and a first-order low-pass filter with an attack time of 5 ms and either a short (40 ms) or a long (2000 ms) release time. As illustrated in the top right panel, the CT was set to 40 dB SPL and a CR of 3:1 was used for all seven frequency channels. Given these fast and slow level estimates, the resulting gain functions, excluding the make-up gain of $M = 10$ dB are shown in the bottom left and right panels, along with the processed output signals at the bottom. It can be seen that the contrast between the peaks and valleys is substantially reduced when a short release time is used, because the level estimate can quickly follow the decrease in the level of the input signal. In contrast, the long release time does not detect soft signal components on a short time scale and only follows the overall signal level. As a consequence, the contrast between soft and loud signal components is not changed, which can lead to under-amplification of low-intensity sounds.

3 Perceptual Effects of Compression

3.1 Audibility

Good speech reception in quiet seems to mainly rely on audibility (French and Steinberg 1947; Fletcher and Galt 1950; Kryter 1962a, b; Pavlovic and Studebaker 1984). Therefore, limited audibility is a crucial factor contributing to speech recognition problems in HI listeners (Souza and Turner 1999; Sherbecoe and Studebaker 2003; Edwards 2004; Souza et al. 2007; Humes and Dubno 2010). It has been proposed that dynamic range compression can provide an improvement over linear amplification in terms of audibility of short, low-energy speech components without compromising loudness comfort (Stelmachowicz et al. 1995; Hickson and Byrne 1997; Souza and Turner 1998; Jenstad and Souza 2005; Alexander and Rallapalli 2017). This, in turn, should translate to an improvement in recognition performance in quiet (Souza et al. 2007). This claim has been supported by results from, e.g., Souza and Turner (1998, 1999) and Davies-Venn et al. (2009), who demonstrated superior audibility due to compression that translated to improved speech recognition performance in HI listeners when the input level of speech was moderate.

In acoustic scenarios with background noise present, speech audibility also appears to be an important contributor to speech recognition. Several studies have compared the performance of HI listeners to NH listeners for whom the hearing loss was simulated by the addition of background noise, resulting in similar audibility.

For example, Zurek and Delhorne (1987) concluded that the loss of audibility is “*the primary source of difficulty in listening in noise for listeners with moderate or milder hearing impairments*”. Furthermore, the results from Desloge et al. (2010) suggest that impaired audibility accounts for the reduced release from masking in fluctuating noise experienced by HI listeners. Similarly, spatial release from masking (SRM) seems to depend strongly on stimulus audibility (Best et al. 2017) and appropriate compensation leads to an improved SRM in HI listeners (Rana and Buchholz 2018). Yet, the relationship between audibility and speech intelligibility of WDRC-amplified speech in noise remains uncertain. It depends on many factors, some of which might also vary with time (Souza et al. 2007). These include: the SNR, temporal characteristics of the noise (e.g., modulation rate and depth, the presence of temporal dips in the noise waveform and their duration), as well as the compression time constants. It has been suggested that the effects of compression would be most prominent in backgrounds that elicit strong fluctuations of the temporal envelope. In such situations, glimpses of the signal can be observed in the temporal dips of the noise. If the system is fast-acting, the applied gain closely follows the fluctuations of the noise, providing increased amplification to the speech glimpses (Gatehouse et al. 2003; Edwards 2004). This effect has been demonstrated, e.g., by Moore et al. (1999) and more recently by Desloge et al. (2017) and Kowalewski et al. (2018). Also, other studies suggested that compression using a short release time and a relatively large number of channels increases the audibility of speech in noise (Moore 2008; Kates 2010; Alexander and Masterson 2014). In contrast, if the compression is too slow, the gain might lag behind the fluctuations in the input signal, resulting in an underamplification of certain speech components (Jerlvall and Lindblad 1978; Stone and Moore 1992; Verschuure et al. 1996; Kuk 1996).

3.2 *Distortions of the Temporal Envelope*

While fast-acting, level-dependent amplification can improve short-term audibility, such processing also changes the internal dynamics of the signal. This includes the reduction in the natural spectral and temporal contrasts (Van Tasell 1993; Gatehouse et al. 2006b), which provides cues for correct speech recognition. The temporal envelope carries important speech information, such as voicing, manner, and prosody (Rosen 1992; Souza and Turner 1996; Davies-Venn and Souza 2014). Preserving the temporal modulation depth, and more specifically, the *modulation spectra*, has been suggested to be crucial for speech recognition (Plomp 1988; Gallun and Souza 2008; Jørgensen and Dau 2011; Zaar and Dau 2016; Alexander and Rallapalli 2017). Plomp (1988) hypothesized that fast-acting compression diminishes the modulation transfer function, leading to reduced intelligibility scores. Results of numerous studies suggest that altering temporal cues may have more pronounced consequences for speech perception of HI compared to NH listeners (Boothroyd et al. 1988; Plomp 1988; Souza and Turner 1996, 1998; van Buuren et al. 1999; Souza and Turner 1999;

Souza and Kitch 2001; Stone and Moore 2003; Souza et al. 2005; Davies-Venn et al. 2009; Souza et al. 2012a, b; Davies-Venn and Souza 2014).

Jenstad and Souza (2005, 2007) and Walaszek (2008) found that temporal envelope fidelity was reduced when using compression with a shorter release time, a higher CR, or a combination of both. This, in turn, led to a decrease in speech recognition performance. In a series of studies, Stone and Moore (2003, 2004, 2007, 2008) showed that multi-channel fast-acting compression not only reduces the modulation depth but also decreases the coherence of modulations across frequency within the speech signal. Moreover, when speech and background noise are processed together in a compressor, they acquire common modulation components which might hinder the target-background separation. Stone and Moore demonstrated that the alterations of the natural temporal structure of the signal mentioned above decrease speech intelligibility. Furthermore, Souza and Gallun (2010) showed that recognition errors of WDRC-amplified consonants could, to a large extent, be accounted for by an alteration of the modulation spectrum of speech, as measured by the spectral correlation index.

The optimal audibility-distortion trade-off remains a topic of ongoing discussion. Villchur (1989) argued that while the reduction in modulation depth does indeed occur with fast-acting compression, the concurrent improvement in audibility may be more important for speech perception in HI listeners than the disrupted temporal cues. On the other hand, Souza et al. (2012a) pointed out that the improvement in audibility was a *necessary but not sufficient* condition for improved speech recognition of WDRC-processed speech. According to Souza et al., the benefit of compression would be observed only if the natural fluctuations of the temporal envelope are preserved. They also suggested that no single recommendation for compression settings should be made. Instead for a given listener and set of acoustic conditions, one should seek a “balance point” where the audibility-distortion trade-off is optimal. Recently, Alexander and Rallapalli (2017) studied the effects of linear amplification, slow- and fast-acting compression on HI listeners’ recognition of fricatives in quiet. They also used various instrumental outcome measures to quantify the effects of compression on the speech dynamic range, audibility, and modulation transfer function. They found that, overall, the fast-acting systems led to an increase in audibility and a decrease of the modulation transfer function. Slow-acting systems, on the other hand, elicited nearly linear behavior and, hence, produced much less detrimental effects on amplitude modulations, at the expense of poorer audibility.

The presence of reverberation might have a complementary effect on the distortion introduced by WDRC. Shi and Doherty (2008) investigated the joint effects of amplification and reverberation on speech recognition and perceived clarity. They considered speech in different reverberation conditions (without noise interferer), combined with linear amplification versus slow- and fast-acting compression. Increasing the reverberation time had a detrimental effect on speech recognition and clarity. WDRC had an overall positive effect on recognition of speech in reverberation, regardless of the compression release time. Slow-, rather than fast-acting compression, however, was rated better for speech clarity. Reinhart et al. (2016) observed that both reverberation and compression distort the temporal envelope. Similarly to the study of

Shi and Doherty (2008), both increasing the amount of reverberation and decreasing the compression release time were detrimental to aided speech recognition. The two effects seem to be additive, as no interaction was observed between compression speed and the amount of reverberation.

3.3 *Effects of Compression on the SNR*

When a speech-in-noise mixture at a positive SNR is processed by a fast-acting compressor, the target signal components become underamplified in relation to the noise, effectively reducing the long-term, broadband SNR at the output (Yund et al. 1987). In contrast, speech components that occur in temporal dips of the noise, or in frequency channels with relatively little noise power, receive a higher gain (relative to the noise) leading to an improvement in the broadband SNR. For speech and noise components falling into the same processing channels, the instantaneous sub-band SNR is not affected by WDRC. The effects discussed in this section are linked to changes in the long-term broadband SNR, unless explicitly stated otherwise. It has been hypothesized by Hagerman and Olofsson (2004) that if speech and the interfering noise have similar spectro-temporal properties, fast-acting compression would reduce the output SNR at positive input SNRs, and improve the output SNR at negative input SNRs, with a “pivotal point” at 0 dB SNR. Furthermore, if the background noise deviates from speech-like spectro-temporal properties, the relationship should remain, but the “pivotal point” would change. An SNR reduction of up to 4 dB was observed in the study of Souza et al. (2006) over a range of input SNRs from -2 to 10 dB. A greater reduction was observed with single-channel rather than two-channel compression. This effect coincided with a deterioration of the temporal envelope fidelity. Naylor and Johannesson (2009) and Rhebergen et al. (2009) conducted systematic studies of the effects of compression parameters on the output SNR and confirmed the previous findings regarding the phenomenon of “SNR compression”. It was also shown that the effect is most prominent for high CRs and short release time constants.

The perceptual consequences of the reduction in SNR due to compression have been studied, e.g., by Souza et al. (2007). It was found that the conditions resulting in an SNR reduction also led to a decrease in speech intelligibility. More recently, Alexander and Masterson (2014) investigated the effects of the compression release time and the number of channels on speech recognition as well as instrumental metrics of output SNR and envelope fidelity. The compression release time interacted with the number of channels in such a way that fewer channels were favorable in a faster compressor, and vice versa. WDRC was found to lead to a reduction of the SNR and more aggressive parameters resulted in a greater reduction, which also coincided with a decreased envelope fidelity. This is in line with studies of Rhebergen et al. (2017) and Desloge et al. (2017). Rhebergen et al. (2017) investigated the acoustical and perceptual effects of conventional WDRC and showed that the output SNR is a reasonably good predictor of the compression benefit for speech in noise. If the output SNR is reduced at the output of the compressor, speech recognition performance will

also decrease. Desloge et al. (2017) proposed a novel energy equalizing (EEQ) signal processing scheme similar to fast-acting compression, which attempts to equalize the power of the signal calculated in short time windows to the estimated long-term power. In fluctuating backgrounds, the system has been shown to enhance the short-term SNR. This SNR enhancement was linked to an improvement in consonant recognition and sentence reception thresholds in HI listeners.

3.4 Binaural Compression

As described above, WDRC affects the variations in signal level across time and frequency, which can constitute important cues for speech intelligibility. Apart from these *monaural cues*, level differences across the two ears (ILDs; Middlebrooks and Green 1991) are utilized for the localization of sounds, while the interaural coherence (IC) and the DRR constitute important cues for certain aspects of spatial perception, such as apparent source width and externalization (Catic et al. 2013; Hassager et al. 2017b). Compression amplification applied independently to the two ear signals can reduce the natural ILDs, potentially affecting sound source localization (Byrne and Noble 1998). Since hearing-aid users' localization abilities are often already affected by hearing loss, it is important to develop compensation strategies that preserve the binaural cues. Keidser et al. (2006) studied the effect of independent binaural WDRC on interaural cues and localization of virtualized sound sources in anechoic conditions in HI listeners. Compression led to a substantial distortion of the ILDs. However, this distortion did not significantly affect listeners' localization of sound sources in the horizontal plane. It is possible that experienced hearing-aid users are able to adapt to the modified spatial cues provided by the hearing-aid processing. On the other hand, Musa-Shufani et al. (2006) showed considerable perceptual consequences of ILD compression for both NH and HI listeners. Compression was found to increase the just noticeable differences (JNDs) of the ILD and to affect the lateralization of high-frequency stimuli. Musa-Shufani et al. reported that the perceived source location was much closer to the mid-line when listening through a binaural WDRC system as compared to a linear reference condition. Similar effects of WDRC in anechoic conditions were reported by Wiggins and Seeber (2011). An increase in apparent source width thereby accompanied the shift in the apparent source position, perception of motion and auditory image splits. Moreover, in a follow-up study (Wiggins and Seeber 2012), the binaurally-compressed stimuli were rated as more diffuse—as opposed to “more focused” with respect to location—and poorly externalized. It has been hypothesized that these effects occur due to relatively slow temporal fluctuations of the ILDs introduced by independent WDRC processing.

To minimize ILD distortion, it has been proposed to link the two compressors. In such a system, the same gain would be applied to the left- and the right-ear signal, removing the ILD fluctuations. A wireless binaural link has already become an industry-standard in state-of-the-art hearing aids and, compared to conventional systems, has been shown to reduce localization errors (Sockalingam et al. 2009).

Moreover, improvements in other outcome measures, such as the rating of naturalness (Sockalingam et al. 2009) or speech intelligibility in relatively spatially-complex scenarios (Kreisman et al. 2010) suggest that the distortions of the spatial scene were decreased with binaurally-linked compression. On the other hand, it has been shown by Hassager et al. (2017b) that certain aspects of spatial perception are still disrupted by WDRC processing despite applying a binaural link. Hassager et al. tested how different types of binaural WDRC affected the abilities of NH and HI listeners to externalize and localize virtualized sound sources in a reverberant environment. They considered four conditions: (i) unprocessed (linear), (ii) independent (unlinked) compression (iii) linked compression and (iv) “ideal” processing, in which the dry signal was compressed prior to convolving it with the binaural room impulse response (BRIR). Both independent and linked binaural compression led to reduced source compactness, an increased diffuseness, image splits and, in some cases, a loss of externalization, as compared to the unprocessed condition. This was due to compression that affects the natural relationship between the direct sound and the reverberant tail. The DRR was decreased by compression, which affected the IC, an important cue for sound externalization (Catic et al. 2013, 2015). Compressing the dry signal maintained the natural DRR and helped to alleviate most of the spatial distortions.

3.5 Towards Scene-Aware Compression

Studies by Yund et al. (1987), Yund and Buckles (1995a), Hornsby and Ricketts (2001), and Rhebergen et al. (2017) reported that the relative speech-intelligibility benefits of different degrees of compression compared to linear amplification depend on factors like the overall input level and the broadband SNR. These results suggest that amplification parameters (e.g., CRs and time constants) should not be chosen with a “one size fits all” concept. Rather, the acoustic scenario should be carefully considered. For example, in a modeling study, Kates (2010) investigated the effects of the number of processing channels and the release time on predicted speech intelligibility and quality. At higher input levels, longer release times resulted in higher intelligibility predictions. At lower intensities, the predicted intelligibility was better in the case of compression than in the case of linear amplification and the effect of the release time was weaker. Based on these results, Kates suggested that compression settings should be adapted according to the listening conditions. Similar ideas were provided by Gatehouse et al. (2003, 2006a), who used the term *auditory ecology* to describe the entirety of conditions that the listeners might encounter and to which the hearing-aid should adapt to.

The signal processing chain of WDRC presented in Sect. 2 is a specific realization of an automatic regulation system, as proposed by Barfod (1978). In his early work on hearing-aid compression, Barfod suggested that such a system could be schematically broken down into several stages, in which signal parameters are measured, the control signal is computed and finally applied to the incoming signal. Barfod acknowledged that the computation stage could follow a simple static rule, as is the

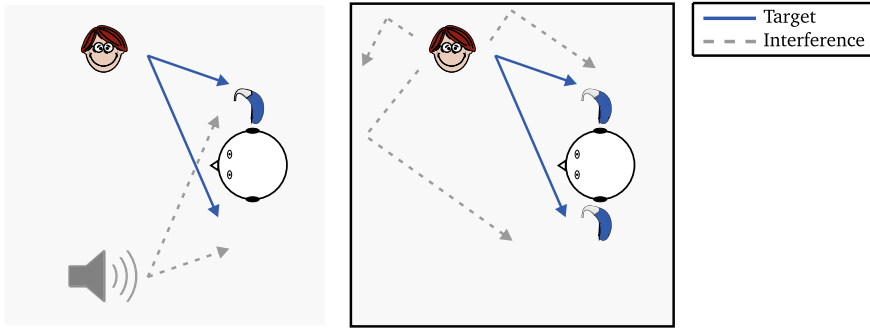


Fig. 3 Application of scene-aware processing, where the target signal is processed with fast-acting compression while the processing of interfering noise (left panel) and room reverberation (middle panel) is effectively linearized via longer time constants

case for conventional WDRC, but the proposed system structure could be a foundation of more complex realizations, in which compression parameters are adjusted in real-time according to some characteristics of the signal. Such adaptive solutions have already become part of commercial hearing aids. They usually adapt the time constants according to the changes in the overall signal level. Examples include the *K-AMP* (Killion et al. 1992), the *dual-front-end automatic gain control* (Moore and Glasberg 1988; Stone et al. 1999) and, more recently, the *guided level estimator* (Neumann 2008), which has been successfully implemented in a commercial product (Simonsen and Behrens 2009). Moreover, Lai et al. (2013) proposed an adaptive WDRC system that adjusted the CR in individual frequency channels depending on the estimated short-term dynamic range. These systems, however, are only sensitive to changes in the overall signal level but do not utilize information related to the presence of the target signal versus the background noise, for example, as reflected by the short-term SNR.

4 Scene-Aware Compression Strategies

As discussed in the previous section, fast-acting WDRC can improve short-term audibility of speech but can also negatively affect the temporal envelope of the signal, introduce undesired across-signal modulations, decrease the output SNR as well as the DRR and introduce ILD fluctuations, thus disrupting spatial cues. Therefore it seems desirable to dynamically change the compression parameters depending on short-term characteristics of the input signal.

The main idea of *scene-aware dynamic range compression* is to adjust key parameters of the WDRC system for a given environment by extracting knowledge about the target signal from the acoustic input. It goes beyond automated program selection (i.e., based on acoustic scene classification), because it utilizes *short-term* estimates

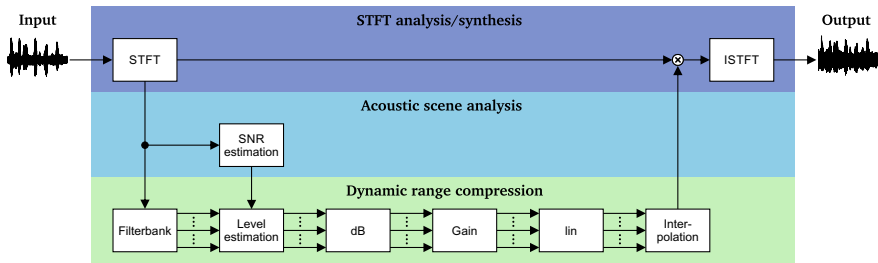


Fig. 4 SNR-aware dynamic range compression consisting of three layers: (1) STFT-based analysis and synthesis stage, (2) short-term SNR estimation and (3) dynamic range compression

reflecting target activity. As illustrated in Fig. 3, the target is assumed to be a speaker, which could be distracted by interfering noise or room reverberation. Thus, the detection of speech activity depends on the acoustic scenario and can be either based on the short-term signal-to-noise ratio (SNR) or the short-term direct-to-reverberant energy ratio (DRR) in different frequency channels. Given an estimation of speech activity, the general idea is to adjust the release time of the compressor for each individual T-F unit. Specifically, fast-acting compression with a short release time is applied to speech-dominated T-F units while the processing of interfering noise or room reverberation is effectively linearized by a longer release time. This target-specific adaptation of the WDRC system aims at combining the advantages of both fast- and slow-acting compression by maximizing the audibility of the target signal while avoiding the majority of artifacts and distortion related to the interference.

In the following, the concept of scene-aware WDRC is applied to two different acoustic scenarios, where the target signal is either processed in the presence of noise or room reverberation. Accordingly, two different estimators for detecting the presence of the target signal are presented, leading to SNR-aware and DRR-aware WDRC systems (May et al. 2018; Hassager et al. 2017a).

4.1 SNR-Aware Dynamic Range Compression

The main building blocks of the SNR-aware WDRC system are illustrated in Fig. 4. Compared to a conventional WDRC system as shown in Fig. 1, the SNR-aware system contains an additional acoustic scene analysis (ASA) layer in which the short-term SNR is estimated for each frequency channel. This short-term SNR is then used to select a short release time of 40 ms for speech-dominated T-F units, while the processing of noise-dominated T-F units is effectively linearized by using a longer release time of 2000 ms.

Short-Term SNR Estimation

Given the STFT representation of the noisy speech signal, the speech power spectral density (PSD) in each individual DFT bin is first obtained using the minimum mean-

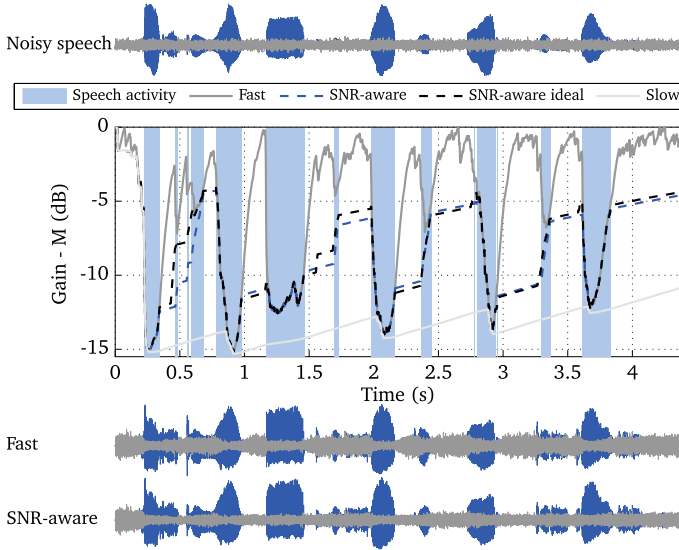


Fig. 5 Speech mixed with ICRA-1 noise at 6 dB SNR (top panel) along with the estimated speech activity and gain functions (excluding the make-up gain M) of four compression systems (fast-acting, slow-acting, SNR-aware and ideal SNR-aware compression) for a frequency channel centered at 2 kHz. The lower two panels show the output of the fast-acting and the SNR-aware compressor, respectively

square error (MMSE) estimator proposed by Erkelens et al. (2007). This MMSE estimator requires knowledge about the noise PSD which is estimated with the noise tracking algorithm proposed by Hendriks et al. (2010). Afterwards, both the PSDs of the noisy speech and the clean speech are passed through the same filterbank employed in the dynamic range compression layer and are subsequently used to estimate the short-term SNR in individual frequency channels (Eaton et al. 2013; May et al. 2017). Finally, the estimated short-term SNR is compared to a predefined threshold in order to detect speech-dominated T-F units (May et al. 2018).

Illustration of SNR-Aware Processing

The principle of SNR-aware dynamic range compression is demonstrated in Fig. 5 for a speech signal mixed with the stationary speech-shaped ICRA-1 noise (Dreschler et al. 2001) at an SNR of 6 dB. Based on the noisy speech signal, the corresponding gain functions of four different approaches are shown for a frequency channel centered at 2 kHz. Conventional fast-acting compression displays fast gain fluctuations. As the level estimator is driven mostly by the speech signal, the gain increases during speech pauses leading to elevation of noise glimpses. This effect can be avoided by using a long release time. However, in such a slow-acting WDRC system, the gain changes slowly and remains relatively low over the entire duration of the stimulus (linearized processing), which might lead to under-amplification of speech components. The SNR-aware system adaptively switches between fast and slow processing

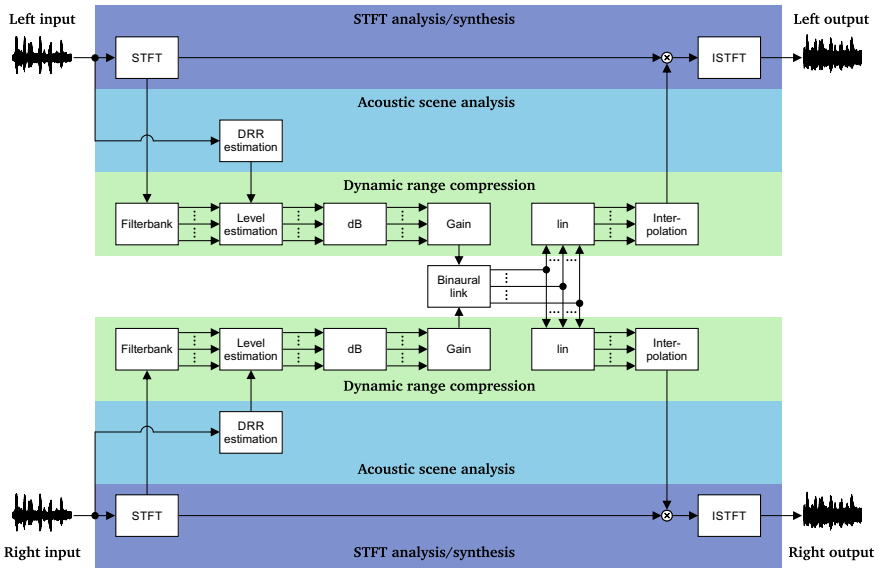


Fig. 6 DRR-aware dynamic range compression consisting of three layers: (1) STFT-based analysis and synthesis stage, (2) short-term DRR estimation and (3) dynamic range compression with a binaural link between both ear signals

depending on the estimated speech activity. Thus, in speech-active time segments, the SNR-aware system can follow rapid intensity changes due to the short release time, while the use of a longer release time for noise-dominated time segments effectively linearizes the processing, which avoids rapid fluctuations in the gain function in response to noise-only segments.

4.2 DRR-Aware Dynamic-Range Compression

The block diagram of the DRR-aware WDRC system is shown in Fig. 6. Similar to the SNR-aware system, information about the presence of the target signal is extracted in the ASA layer. Specifically, a monaural variance-based estimator is used to detect T-F units that are dominated by the direct sound. This classification is subsequently used to selectively apply fast-acting compression to T-F units dominated by the direct sound, while reverberation-dominated T-F units are processed with slow-acting compression. Although the gains were computed independently for the left and the right ear signals, the final gain was linked by taking the minima of the left and right gain values to preserve the natural ILDs.

Short-Term DRR Estimation

The input signal is passed through a bank of seven octave-spaced band-pass filters that matched the frequency resolution of the dynamic range compression layer. Then, a variance-based feature was computed across short time windows of 10-ms duration with 75 % overlap (Hazrati et al. 2013). The rationale behind this monaural estimator proposed by Hazrati et al. (2013) is that the variance-based feature is higher for signal components dominated by the direct sound and reduces if the signal is dominated by reverberation. An adaptive threshold is used to incorporate some temporal context and to ensure that the decision threshold is adjusted to the overall feature level in a given acoustic condition. This was found to be important when dealing with different rooms and sound sources directions. More details concerning the detection of direct-sound components can be found in Hassager et al. (2017a).

5 Instrumental Metrics

A wide range of instrumental metrics have been proposed to analyze the acoustical signal at the output of compression systems. Many of them are good predictors of recognition of compressed speech or the perceived sound quality. Some of the perceptually-relevant metrics are discussed below.

5.1 Separation of Speech and Noise Components

In order to compute objective metrics, such as the SNR at the output of a compression system, special techniques are required to separate the influence of the signal-dependent and time-varying gain function on the speech and the noise components.

The phase-inversion technique proposed by Hagerman and Olofsson (2004) is probably the most widely-used approach. In this method, two mixtures are generated, with identical target speech but noise in positive and negative polarity. Afterwards, both mixtures are processed by the nonlinear system and the speech- and noise-alone signals can be estimated by adding or subtracting both processed mixtures. This technique assumes that the nonlinear terms introduced by the system are negligible. However, it was shown by Rhebergen (2006) and Rhebergen et al. (2008) that this assumption is often invalid, leading to estimation errors. Nonetheless, the phase inversion technique is still the only available solution if the system under investigation is a “black box”.

Alternatively, if details about the gain function are available, one can employ a technique known as *shadow filtering* (Gustafsson et al. 1996; Fredelake et al. 2012). This technique has been successfully utilized, for example, by Rhebergen et al. (2009), Kowalewski et al. (2018) and May et al. (2018). The general principle of

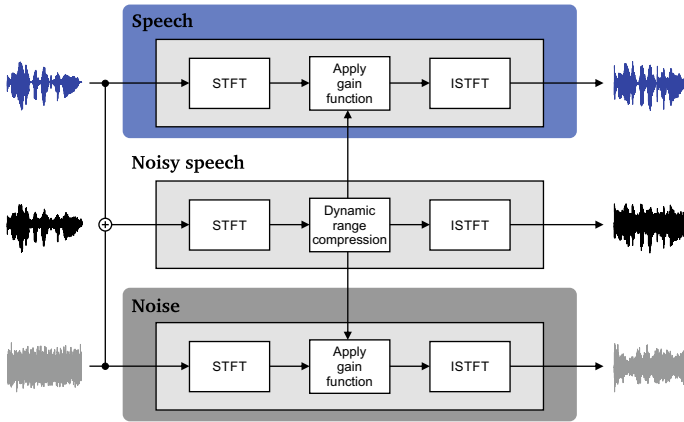


Fig. 7 Principle of shadow filtering: The gain function of the compressor is estimated based on the noisy speech signal and then applied separately to the clean speech, the noise, and the noisy speech signals

shadow filtering is illustrated in Fig. 7, where the signal-dependent gain function of the compressor is estimated based on the noisy speech signal. In order to investigate the impact on speech and noise components, the same gain function can be applied separately to the speech and the noise signals. This method will be used in Sect. 6.1 to analyze how a set of objective metrics reflect changes in speech, noise, and noisy speech before and after processing.

5.2 Broadband Input/Output SNR Analysis

As mentioned in Sect. 3.3, the long-term output SNR is a valuable metric for predicting intelligibility of WDRC-processed speech (Rhebergen et al. 2009). Depending on the characteristics of the speech and the interferer as well as their relative input levels, compression can either increase or decrease the output SNR (Souza et al. 2006; Naylor and Johannesson 2009). To investigate these effects, noisy speech mixtures can be generated and processed at various input SNRs and shadow-filtering can be used to calculate the long-term SNR before and after processing. An example of such an “input/output” function is shown in Fig. 10 for four different compression systems.

5.3 Effective Compression Ratio

The degree to which compression affects the distribution of the short-term signal level depends on many factors beyond the nominal CR. These include the compression time constants as well as the characteristics of the input signal, such as the amount of background fluctuations and the input SNR. The effective compression ratio (ECR)

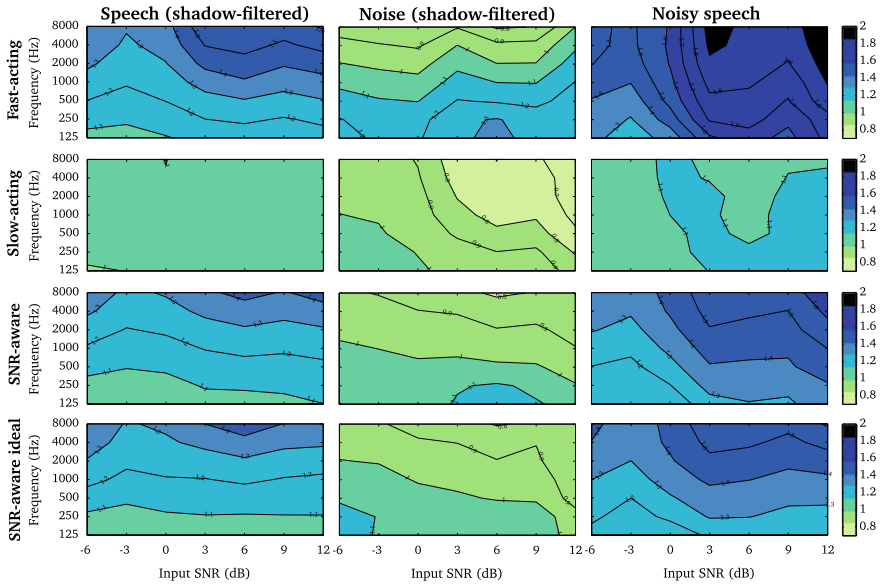


Fig. 8 Contours of effective compression ratios (ECRs) for the fast-acting (first row), slow-acting (second row), SNR-aware (third row), and ideal SNR-aware compressors (fourth row) as a function of the input SNR and the channel center frequency. Results were averaged across all four noise types, which are described in Sect. 6.1. The left, middle, and right columns show results for shadow-filtered speech, shadow-filtered noise, and noisy speech

can be defined as the ratio of the dynamic range at the input and the output of a WDRC system (Souza et al. 2006; Croghan et al. 2014). Previous literature focused on the dynamic range analysis of compressed speech in quiet. Hence, the upper limit used to define the dynamic range was typically based on the 98th or the 99th percentile of the short-term level distribution, while the lower limit could be as low as the 5th percentile. However, when considering noisy speech, it might be possible that HI listeners are not able to utilize such low dips in the signal. Therefore, the dynamic range (and the resulting ECRs) can alternatively be defined as the difference between the 99th and the 50th percentiles. This analysis can be performed in different frequency bands and across a wide range of input and output SNRs. With the use of shadow-filtering, the effective compression of speech and noise components can be studied in isolation, as shown in Fig. 8.

5.4 Modulation-Spectrum Analysis

As mentioned in Sect. 3.2, preserving the modulation spectrum of speech is important for speech recognition. Moreover, introducing additional modulations of the background could lead to an overall decrease in the SNR—affecting both quality and intelligibility. It is therefore valuable to analyze the effects of WDRC processing

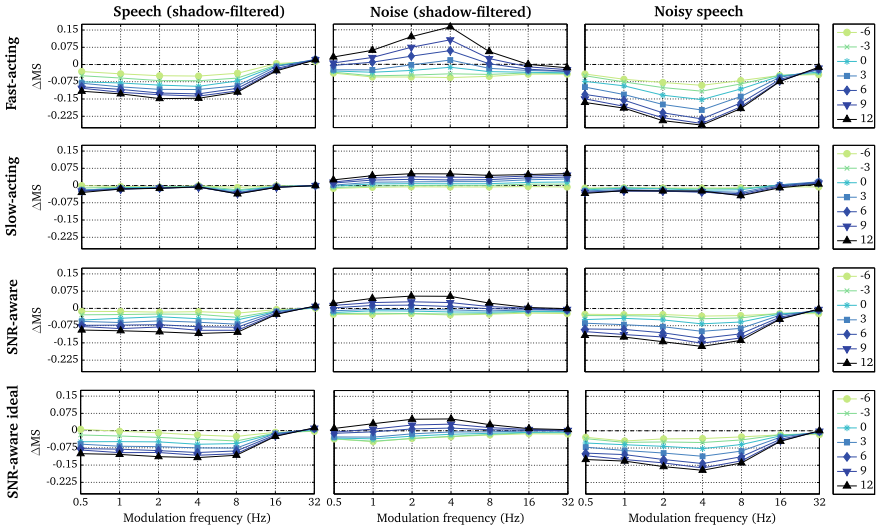
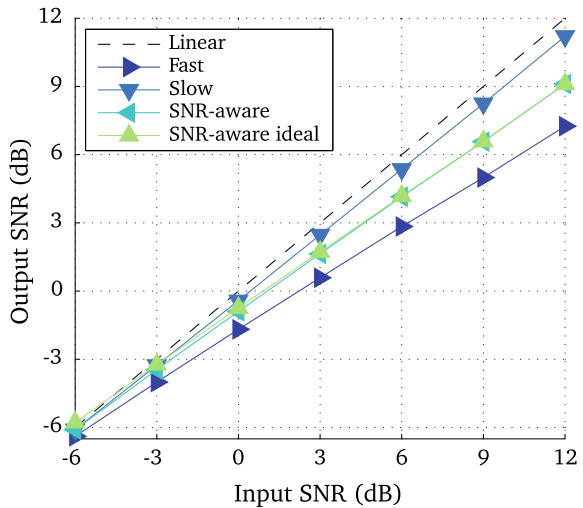


Fig. 9 Relative change in the modulation spectrum (ΔMS) due to fast-acting (first row), slow-acting (second row), SNR-aware (third row), and ideal SNR-aware compressors (fourth row) as a function of the modulation frequency and input SNR. Results were averaged across all four noise types, which are described in Sect. 6.1. The black dashed line indicates the zero line while the left, middle, and right columns show results for shadow-filtered speech, shadow-filtered noise, and noisy speech

Fig. 10 Input/output SNR of four different WDRC systems (fast-acting, slow-acting, SNR-aware and ideal SNR-aware compression) and a linear reference. Results were averaged across all four noise types, which are described in Sect. 6.1



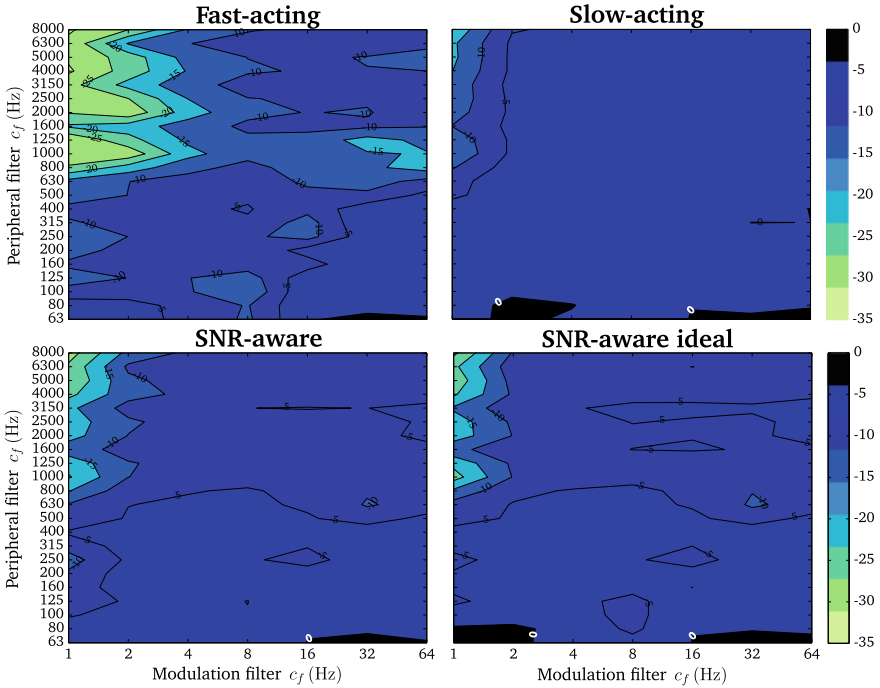


Fig. 11 $\Delta\text{SNR}_{\text{env}}$ metric in dB for the fast-acting (top left), slow-acting (top right), SNR-aware (bottom left) and ideal SNR-aware compressor (bottom right) as a function of the peripheral and the modulation center frequency. Results were averaged across all four noise types, which are described in Sect. 6.1

on the modulation spectra of the compressed speech, the noise and the noisy speech mixture. They can be obtained with the aid of shadow-filtering, by computing the DFT of the Hilbert envelope of each signal and relating the root mean square (RMS) in octave-spaced modulation-frequency bands to the DC component of the envelope, ranging from 1 to 32 Hz (Dreschler et al. 2001). The relative change in the modulation spectrum due to compression (ΔMS) can be calculated by comparing the modulation spectrum before and after processing. Examples of ΔMS patterns are shown in Fig. 9.

5.5 SNR in the Envelope Domain

Changes in the modulation spectrum of speech and noise may affect the broadband SNR as well as the SNR in the envelope domain, as discussed in Sect. 3.2. Ewert and Dau (2000) proposed a model which accounts for modulation detection and masking data using the ratio of target-to-masker power at the output of a modulation filterbank (SNR_{env}). The metric has been successfully used to predict speech recognition perfor-

mance of NH listeners in a variety of processing conditions, including the influence of background noise, the presence of reverberation, spectral subtraction and phase jitter that modified the modulation content of the stimuli (Jørgensen and Dau 2011; Jørgensen et al. 2013). The calculation of the SNR_{env} requires a priori knowledge of the speech and noise signals. Both signals are processed through a bank of peripheral filters. Then, the Hilbert envelope is extracted in each peripheral channel and processed by a modulation filterbank. The ratio of the envelope power of the signal and the noise at the output of each modulation filter is calculated, yielding the SNR_{env} . The effects of compression can be evaluated by computing a difference in the SNR_{env} between the processed and unprocessed signals. Examples of $\Delta\text{SNR}_{\text{env}}$ across peripheral channel frequencies and modulation channel frequencies are shown in Fig. 11.

5.6 Analysis of Spatial Cues

Hassager et al. (2017b) proposed a set of instrumental metrics to analyze the effect of compression on spatial perception in a reverberant environment. These metrics included the ILD distribution, the IC, and the DRR. The distribution of the ILDs was calculated in auditory subbands by analyzing the input signal with a gammatone filterbank. The envelopes of the filterbank outputs were segmented using a rectangular window of 20-ms duration with 50 % overlap. The power in each window was converted to decibels (dB SPL), and the differences between the left- and right-ear levels were then used to compute an ILD histogram. The IC was computed in each subband using the normalized cross-correlation between the left- and right-ear signals, considering lag values, τ , between -1 and 1 ms. The maximum absolute value of the cross-correlation was found across the lag value dimension, τ , and subsequently used to create an IC histogram. Finally, the DRR was calculated in the frequency domain. The direct part was defined as the part of the signal occurring prior to the first reflection (2.5 ms in case of Hassager et al. 2017b) and the later part was treated as reverberation. The PSDs of the direct and reverberant parts were normalized by the power of the corresponding dry signal. Subsequently, a ratio of the normalized PSDs was converted to decibels.

Measuring the shift in the ILD distribution can be used to predict potentially disruptive effects of WDRC on sound-source localization abilities, as discussed in Sect. 3.4. Moreover, if substantial changes in the DRR and the IC are observed, this can be indicative of potential distortions of spatial perception, such as the increase in apparent source width, the occurrence of image splits or the loss of externalization.

6 Results Obtained with Scene-Aware Compression Systems

In this section, results from two studies on scene-aware dynamic range compression are summarized. May et al. (2018) conducted an instrumental evaluation of conventional fast-acting, slow-acting and SNR-aware compression. This evaluation is presented in Sect. 6.1 and includes the analysis of ECRs, changes in the modulation spectrum and the input/output SNR as described in Sect. 5. Moreover, Hassager et al. (2017a) evaluated the spatial perception of NH listeners with conventional fast-acting, slow-acting, and DRR-aware compression. In Sect. 6.2, the listeners' responses in terms of the perceived source position and apparent source width are presented.

6.1 Speech in Noise

Stimuli

Noisy speech sampled at a rate of 16 kHz was created by mixing clean speech from the Danish hearing in noise test (HINT) corpus with four different types of background noise at seven broadband input SNRs: -6 , -3 , 0 , 3 , 6 , 9 and 12 dB. The following noise types were considered: the stationary ICRA-1 noise and the non-stationary ICRA-7 noise based on a six-talker babble (Dreschler et al. 2001), as well as car noise and factory noise from the NOISEX database (Varga and Steeneken 1993). Following Naylor and Johannesson (2009), all noise types were spectrally matched to the long-term average spectrum of the Danish HINT corpus.

Compression Parameters

The CTs of seven octave-wide frequency channels were calibrated using a stationary noise that was spectrally matched to the long-term average spectrum of the HINT speech material. The nominal CRs and the make-up gain were derived from the NAL-NL2 (Keidser et al. 2011) gain prescription for the N_4 standard audiogram (sloping, moderate-to-severe hearing loss following Bisgaard et al. (2010)) using the settings *slow* and *unilateral*. All parameters are summarized in Table 1. The nominal attack time was set to 5 ms in all cases while the fast-acting and slow-acting compressors used nominal release times of 40 and 2000 ms. As explained in Sect. 4.1, the SNR-aware approach used an estimation of the short-term SNR to switch between fast- and slow-acting compression depending on whether individual T-F units were dominated by speech (high SNR) or background noise (low SNR). The ideal SNR-aware system used the a priori SNR, which was calculated from the individual speech and noise signals.

Results

Figure 8 shows the ECRs obtained with conventional fast-acting (first row), slow-acting (second row), SNR-aware (third row) and ideal SNR-aware compression (fourth row). The ECRs based on the shadow-filtered speech, shadow-filtered noise and the noisy speech mixture, are shown in the left, middle, and right columns, respectively. Within each panel, the ECRs are shown as a function of frequency and input SNR. Overall, the ECRs were significantly lower than the nominal ratios shown in Table 1. However, the highest compression ratios were still observed in the high-frequency channels, reflecting the frequency-dependent characteristic of the ratios prescribed by the NAL-NL2 procedure. The ECRs for the shadow-filtered speech obtained with the conventional fast-acting system ranged from 1.1 to 1.6. The noise signal was less compressed than speech, with the highest ECR of 1.3 occurring in the low-frequency channels. In the higher frequency channels, the noise became effectively expanded, as the computed ratios were below 1.0. The ECRs obtained for speech with slow-acting compression were close to 1.0, regardless of the input SNR and frequency. The noise was also not compressed, with ratios smaller or equal to 1.0. The contour plots for the speech processed with SNR-aware and ideal SNR-aware systems were very similar to each other, with ECRs ranging from 1.1 to 1.4. Small amounts of compression were observed in the noise signal (ECRs up to 1.1), but again in most channels, the effective ratios were smaller or equal to 1.0.

Figure 9 shows the relative change in the modulation spectrum (ΔMS , see Sect. 5.4 for more details) for the different signals and processing types. Within each panel, the ΔMS values are plotted as a function of the modulation frequency and color coded according to the input SNR. Negative values indicate a reduction, while positive values indicate an increase in the modulation depth. All processing types led to some degree of reduction of speech modulations, at least for the modulation frequencies from 0.5 to 16 Hz. The amount of reduction increased (more negative ΔMS) with increasing SNR. The noise modulation depth was also decreased at smaller SNRs but as the SNR became larger, an enhancement was observed. Conventional fast-acting compression led to a maximum reduction of speech modulations of around 0.15 at the SNR of 12 dB. At the same time, noise modulations were enhanced with ΔMS up to 0.15 centered around a prominent 4-Hz peak. The changes in the modulation spectrum introduced by the conventional slow-acting compression were much smaller in magnitude. The values of ΔMS for the shadow-filtered speech signal were between 0 and -0.075 and were less dependent on the input SNR. For the noise

Table 1 Compression thresholds (CTs) in decibels, nominal compression ratios (CRs) and make-up gain M in decibels for individual channel center frequencies used by the SNR-aware compression system

	Channel center frequency (Hz)						
	125	250	500	1000	2000	4000	8000
CT (dB)	43	43	41	41	37	31	28
CR	2.2:1	2.2:1	2.2:1	3.0:1	3.5:1	3.3:1	2.5:1
M (dB)	22.1	22.1	24.4	34.5	39.4	43.5	42.5

signal, the increase in ΔMS did not exceed 0.075, and the pattern was overall flatter than for fast-acting compression. The two SNR-aware systems exhibited similar ΔMS -patterns. The patterns of modulation-depth reduction for speech resembled those observed with the conventional fast-acting system, but smaller in magnitude (the maximum reduction lied between 0.075 and 0.15). The patterns for noise, on the other hand, more closely resembled slow-acting compression, with an enhancement that did not exceed the value of 0.075 and a lack of a prominent peak.

The input/output SNR functions for each system are shown in Fig. 10 (see Sect. 5.2 for more details). The dashed line indicates the linear reference. All compression systems led to a varying degree of SNR reduction, which became more pronounced at higher input SNRs. The greatest amount of SNR reduction, up to 4.8 dB, was observed for the conventional fast-acting compression. The slow-acting system was much closer to linear, with the reduction of the SNR not exceeding 2 dB. Both SNR-aware systems introduced a similar amount of reduction, which did not exceed 3 dB.

Figure 11 shows the relative change in the SNR_{env} metric in dB before and after processing (see Sect. 5.5 for more details) as a function of the peripheral and modulation center frequency for the four different compression systems. A reduction in the SNR_{env} was observed for all four systems. The greatest amount of reduction was introduced by fast-acting compression, with up to 25 dB in the peripheral channels above 1000 Hz and for low modulation frequencies. Slow-acting compression led to a less drastic change in the SNR_{env} . The maximum reduction was 20 dB but this occurred only around 6000 to 8000 Hz and in the lowest modulation filter. The ΔSNR_{env} metric of both SNR-aware systems elicited similar patterns and the amount of SNR reduction was much more similar to the one obtained with the slow-acting system.

6.2 *Speech in a Reverberant Environment*

Stimuli

Clean speech from the Danish HINT corpus was sampled at a rate of 48 kHz and convolved with BRIRs. Individual BRIRs for each listener were measured using a maximum length sequence (MLS) signal played from the loudspeaker at 300° azimuth (equivalent to 60° to the right with respect to the frontal direction, see Fig. 12) and recorded using two DPA high sensitivity microphones placed at the ear-canal entrances.

Compression Parameters

The compression system operated in seven octave-wide frequency channels with center frequencies spaced between 125 and 8000 Hz. The nominal CRs ranged from 3.4 to 4.0. The attack and release times were 10 and 60 ms in the *fast* mode and both 2000 ms in the *slow* mode (time constants defined according to ANSI S3.22-1996). As discussed in Sect. 4.2, the DRR-aware WDR system switched between

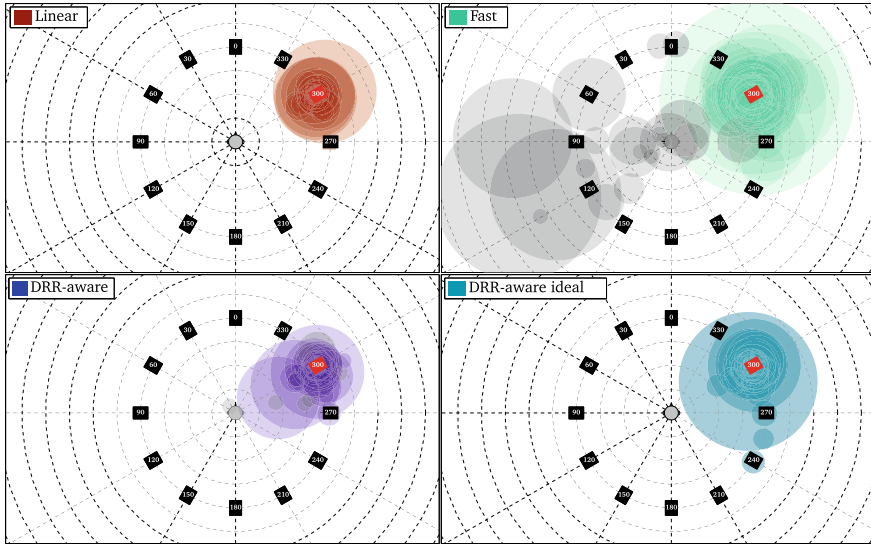


Fig. 12 Listeners’ responses for the four tested compression systems: linear (top left), fast-acting (top right), DRR-aware (bottom left) and ideal DRR-aware compression (bottom right). The response of each listener is indicated as a transparent filled circle with a center and width corresponding to the associated perceived sound image. The main sound images are shown by the different colors while image splits are indicated in gray

fast- and slow-acting compression depending on whether individual T-F units were dominated by the direct sound (high DRR) or room reverberation (low DRR). The ideal DRR-aware system used a priori information about the BRIRs to detect direct sound activity.

Listeners and Experimental Setup

Eighteen NH listeners aged between 19 and 35 years participated in the study. The experiment took place in a reverberant listening room designed in accordance with the IEC 60268-13 (1985) standard. The reverberation time T_{30} was approximately 500 ms, representing a typical living room environment. The listeners were seated in a chair equipped with a headrest and a response touchscreen. Twelve Dynaudio BM6 loudspeakers surrounded the listening position in a circular arrangement with a radius of 150 cm. A graphical representation of the loudspeaker arrangement was displayed on the touchscreen. The NH listeners were asked to place circles on the screen, according to the perceived position and width of the sound sources. Multiple circles could be positioned, in case of image splits were perceived. The test sound was always presented from the loudspeaker at 300° azimuth (equivalent to 60° to the right with respect to the frontal direction, see Fig. 12).

Results

The graphical representations of the listeners' responses are shown in Fig. 12. All responses are overlaid in each panel. The circles indicate the position and width of the perceived sources, while gray circles indicate image splits. The top left panel shows the responses in the linear (unprocessed) condition. The perceived sources were externalized, correctly localized and relatively compact, with no image splits. The right top panel shows the responses obtained with conventional binaurally-linked fast-acting compression. For several listeners, the perception of the source position deviated from the target. Many image splits were reported, and some listeners perceived internalized sources moving between the ears. The apparent source width was also far greater than it was the case in the linear condition. The responses for the linked ideal DRR-aware system based on a priori knowledge are shown in the bottom right panel. There were still a few cases of incorrect localization. However, compared to the conventional fast-acting compression system, the perceived sources were much more compact, and there were no image splits. The response pattern for the linked DRR-aware system (bottom left panel) was similar, but a limited number of image splits were reported.

7 Discussion and Conclusion

This chapter provided an overview of hearing-aid compression strategies. A novel scene-aware dynamic range compression strategy was presented that adjusts the release time of the compressor depending on short-term estimates reflecting target activity. The proposed strategy aims at combining the advantages of both fast- and slow-acting compression and has been evaluated for two acoustic scenarios reflecting speech in noise and speech in a reverberant environment. The SNR-aware strategy has been evaluated in terms of instrumental metrics and compared to conventional fast- and slow-acting compression. Consistent with previous studies (e.g., Alexander and Rallapalli 2017), it has been demonstrated that fast-acting compression introduces a relatively high effective compression of the speech signal. However, at the same time, it significantly reduces the speech modulation depth and introduces modulations to the background noise. This effect is most pronounced at high input SNRs, where fluctuations of the speech-signal envelope drive the compression gain. The peaks of the speech signal become flattened while the compressor gain rapidly increases in the pauses, cyclically amplifying portions of the noise. This results in a pronounced reduction in the energetic, broadband SNR, as well as in the SNR in the envelope domain (SNR_{env}). The reduction of the broadband SNR may be linked to decreased speech audibility and recognition, as suggested by recent studies of Rhebergen et al. (2017) and Alexander and Rallapalli (2017). The decreased SNR_{env} might indicate that speech envelope cues are degraded and that the target and the background attain common components in the modulation domain after applying fast-acting compression. This makes them perceptually more similar and more difficult to separate.

Moreover, fast-acting compression may lead to a sensation of “pumping” and an increased perceived noisiness (e.g., Kuk 1996; Neuman et al. 1998; Kates and Arehart 2014).

The use of slow-acting compression helps avoid most of those distortions. As seen in Fig. 8, slow-acting compression behaves almost linearly in terms of the input/output SNR. It also does not cause drastic changes in the modulation spectrum (ΔMS) of speech and noise and introduces only a minimal reduction of the SNR_{env} . However, the ECRs obtained with slow-acting compression are close to 1.0, which indicates that the system behaves essentially linearly, i.e., it does not provide compression of the speech target and is, therefore, less effective in restoring audibility.

The proposed SNR-aware compression strategy has similar effects on the speech signal as the conventional fast-acting system in terms of ECR and ΔMS . The reduction of the speech modulation depth is unavoidable if an effective compression of the signal is desired. However, the SNR-aware system minimizes the interaction between speech and noise within the compressor, improving the noise modulation fidelity and providing better broadband SNR and SNR_{env} compared to the conventional fast-acting strategy. The perceptual effect of the SNR-aware systems are yet to be evaluated. The potential perceptual assessment would include not only speech intelligibility, but also aspects related to sound quality and cognitive effort.

In terms of spatial perception, conventional fast-acting compression has been shown to introduce substantial distortions of the acoustic scene. Even though the use of a binaural link avoided the ILD fluctuations, the fast gain fluctuations led to a disruption of the DRR, which resulted in a loss of externalization, image splits, and an increased apparent source width. The DRR-aware compression system seems to avoid most of the spatial distortions introduced by conventional fast-acting compression. It improves listeners’ spatial perception, which was evaluated in aided NH listeners. It is likely that the improvements in the spatial fidelity will not only improve the perceived quality of the acoustic scene, but that they may also enable better source separation. This, in turn, might be manifested in improved intelligibility of speech in noise and a lower cognitive effort required for speech understanding.

The accuracy of the SNR and DRR estimators could be improved by the use of supervised-learning techniques, which have been utilized, for example, by Wang and Chen (2018) and May (2018). Using a joint estimator—e.g., such as the one proposed by Kuklasinski et al. (2016)—the DRR-aware and SNR-aware approaches could be combined and applied in complex scenarios where both background noise and room reverberation are present simultaneously. However, many challenging acoustical scenarios involve speech-on-speech masking. In such conditions, conventional SNR-based estimators would likely fail to reliably detect the presence of the target speaker, rendering SNR-aware compression impractical. In such conditions, spatial cues could be exploited to identify individual T-F units dominated by the target. Then, similar principles as employed by the SNR-aware system could be applied, namely fast-acting compression of the target and linearization of the interfering sources. For example, the spatial separation of sound sources could be exploited by using adaptive spatial filtering (Doclo et al. 2015) or robust spatial localization techniques (May et al. 2013; Ma et al. 2017). Thanks to the recent technological developments, such

spatial filters could be additionally driven by eye-gaze (Favre-Félix et al. 2017) or attention (Wong et al. 2018). These developments open new possibilities for *target-aware* compression.

Acknowledgements The authors are indebted to two anonymous reviewers for remarks that helped improve the manuscript.

References

- Alexander, J.M., and Masterson, K. 2014. Effects of WDRC release time and number of channels on output SNR and speech recognition. *Ear and Hearing* 1–15.
- Alexander, J.M., and V. Rallapalli. 2017. Acoustic and perceptual effects of amplitude and frequency compression on high-frequency speech. *Journal of the Acoustical Society of America* 142 (2): 908–923.
- Allen, J.B. 1977. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Transactions on Audio, Speech, and Signal Processing* 25(3), 235–238.
- ANSI S3.22-1996. 1996. *Specification of Hearing Aid Characteristics*. American National Standards Institute.
- ANSI S3.5. 1997. *Methods for the Calculation of the Speech Intelligibility Index*. American National Standards Institute.
- Barfod, J. 1978. Automatic regulation systems with relevance to hearing aids. *Scandinavian Audiology Supplementum* 6: 355–378.
- Best, V., C.R. Mason, J. Swaminathan, E. Roverud, and G. Kidd Jr. 2017. Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures. *Journal of the Acoustical Society of America* 141 (1): 81–91.
- Bisgaard, N., M.S. Vlaming, and M. Dahlquist. 2010. Standard audiograms for the IEC 60118–15 measurement procedure. *Trends in Amplification* 14 (2): 113–120.
- Boothroyd, A., N. Springer, L. Smith, and J. Schulman. 1988. Amplitude compression and profound hearing loss. *Journal of Speech, Language and Hearing Research* 31 (3): 362–376.
- Bor, S., P. Souza, and R. Wright. 2008. Multichannel compression: Effects of reduced spectral contrast on vowel identification. *Journal of Speech, Language and Hearing Research* 51 (5): 1315–1327.
- Bustamante, D.K., and L.D. Braida. 1987. Multiband compression limiting for hearing-impaired listeners. *Journal of Rehabilitation Research and Development* 24 (4): 149–160.
- Byrne, D., and W. Noble. 1998. Optimizing sound localization with hearing aids. *Trends in Amplification* 3 (2): 51–73.
- Catic, J., S. Santurette, J.M. Buchholz, F. Gran, and T. Dau. 2013. The effect of interaural-level-difference fluctuations on the externalization of sound. *Journal of the Acoustical Society of America* 134 (2): 1232–1241.
- Catic, J., S. Santurette, and T. Dau. 2015. The role of reverberation-related binaural cues in the externalization of speech. *Journal of the Acoustical Society of America* 138 (2): 1154–1167.
- Cox, R.M., J.A. Johnson, and J. Xu. 2016. Impact of hearing aid technology on outcomes in daily life I: The patients' perspective. *Ear and Hearing* 37 (4): e224–e237.
- Croghan, N.B., K.H. Arehart, and J.M. Kates. 2014. Music preferences with hearing aids: Effects of signal properties, compression settings, and listener characteristics. *Ear and Hearing* 35 (5): e170–e184.
- Davies-Venn, E., and P. Souza. 2014. The role of spectral resolution, working memory, and audibility in explaining variance in susceptibility to temporal envelope distortion. *Journal of the American Academy of Audiology* 25 (6): 592–604.

- Davies-Venn, E., P. Souza, M. Brennan, and G.C. Stecker. 2009. Effects of audibility and multichannel wide dynamic range compression on consonant recognition for listeners with severe hearing loss. *Ear and Hearing* 30 (5): 494.
- De Gennaro, S., L. Braidà, and N. Durlach. 1986. Multichannel syllabic compression for severely impaired listeners. *Journal of Rehabilitation Research and Development* 23 (1): 17–24.
- Desloge, J.G., C.M. Reed, L.D. Braidà, Z.D. Perez, and L.A. D'Aquila. 2017. Masking release for hearing-impaired listeners: The effect of increased audibility through reduction of amplitude variability. *Journal of the Acoustical Society of America* 141 (6): 4452–4465.
- Desloge, J.G., C.M. Reed, L.D. Braidà, Z.D. Perez, and L.A. Delhorne. 2010. Speech reception by listeners with real and simulated hearing impairment: Effects of continuous and interrupted noise. *Journal of the Acoustical Society of America* 128 (1): 342–359.
- Dillon, H. 1999. NAL-NL1: A new procedure for fitting non-linear hearing aids. *Hearing Journal* 52 (4): 10–16.
- Dillon, H. 2008. *Hearing Aids*. Hodder Arnold.
- Dillon, H., and L. Storey. 1998. The national acoustic laboratories' procedure for selecting the saturation sound pressure level of hearing aids: Theoretical derivation. *Ear and Hearing* 19 (4): 255–266.
- Doclo, S., W. Kellermann, S. Makino, and S.E. Nordholm. 2015. Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones. *IEEE Signal Processing Magazine* 32 (2): 18–30.
- Dreschler, W.A., H. Verschuure, C. Ludvigsen, and S. Westermann. 2001. ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. *International Journal of Audiology* 40 (3): 148–157.
- Eaton, J., M. Brookes, and P.A. Naylor. 2013. A comparison of non-intrusive SNR estimation algorithms and the use of mapping functions. In *Proceedings of the EUSIPCO*, 3–7.
- Edwards, B. 2004. Hearing aids and hearing impairment in Speech processing in the auditory system. In *Speech Processing in the Auditory System, Chap. 7*, ed. S. Greenberg, W.A. Ainsworth, A.N. Popper, and R.R. Fay, 339–421. New York, NY: Springer.
- Erkelens, J.S., R.C. Hendriks, R. Heusdens, and J. Jensen. 2007. Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors. *IEEE Transactions on Audio, Speech and Language* 15 (6): 1741–1752.
- Ewert, S.D., and T. Dau. 2000. Characterizing frequency selectivity for envelope fluctuations. *Journal of the Acoustical Society of America* 108 (3): 1181–1196.
- Favre-Félix, A., R. Hietkamp, C. Graversen, T. Dau, and T. Lunner. 2017. Steering of audio input in hearing aids by eye gaze through electrooculography. In *Proceedings of the International Symposium on Auditory Audiology Research (ISAAR)*, vol. 6, 135–142.
- Fletcher, H., and R.H. Galt. 1950. The perception of speech and its relation to telephony. *Journal of the Acoustical Society of America* 22 (2): 89–151.
- Fredelake, S., I. Holube, A. Schlueter, and M. Hansen. 2012. Measurement and prediction of the acceptable noise level for single-microphone noise reduction algorithms. *International Journal of Audiology* 51 (4): 299–308.
- French, N.R., and J.C. Steinberg. 1947. Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America* 19 (1): 90–119.
- Gallun, F., and P. Souza. 2008. Exploring the role of the modulation spectrum in phoneme recognition. *Ear and Hearing* 29 (5): 800–813.
- Gatehouse, S., G. Naylor, and C. Elberling. 2003. Benefits from hearing aids in relation to the interaction between the user and the environment. *International Journal of Audiology* 42 (sup1): 77–85.
- Gatehouse, S., G. Naylor, and C. Elberling. 2006a. Linear and nonlinear hearing aid fittings—1. Patterns of benefit. *International Journal of Audiology* 45 (3): 130–152.
- Gatehouse, S., G. Naylor, and C. Elberling. 2006b. Linear and nonlinear hearing aid fittings—2. Patterns of candidature. *International Journal of Audiology* 45 (3): 153–171.

- Giannoulis, D., M. Massberg, and J.D. Reiss. 2012. Digital dynamic range compressor design: A tutorial and analysis. *Journal of Audio Engineering Society* 60 (6): 399–408.
- Grimm, G., T. Herzke, D. Berg, and V. Hohmann. 2006. The master hearing aid: A PC-based platform for algorithm development and evaluation. *Acta Acustica United with Acustica* 92 (4): 618–628.
- Gustafsson, S., R. Martin, and P. Vary. 1996. On the optimization of speech enhancement systems using instrumental measures. In *Workshop on Quality Assessment in Speech*, 36–40. Audio and Image Communication.
- Hagerman, B., and Å. Olofsson. 2004. A method to measure the effect of noise reduction algorithms using simultaneous speech and noise. *Acta Acustica United with Acustica* 90 (2): 356–361.
- Hassager, H.G., T. May, A. Wiinberg, and T. Dau. 2017a. Preserving spatial perception in rooms using direct-sound driven dynamic range compression. *Journal of the Acoustical Society of America* 141 (6): 4556–4566.
- Hassager, H.G., A. Wiinberg, and T. Dau. 2017b. Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment. *Journal of the Acoustical Society of America* 141 (4): 2556–2568.
- Hazrati, O., L. Jaewook, and P.C. Loizou. 2013. Blind binary masking for reverberation suppression in cochlear implants. *Journal of the Acoustical Society of America* 133 (3): 1607–1614.
- Hendriks, R.C., R. Heusdens, and J. Jensen. 2010. MMSE-based noise PSD tracking with low complexity. In *Proceedings of the ICASSP*, 4266–4269.
- Henning, R.L.W., and R.A. Bentler. 2008. The effects of hearing aid compression parameters on the short-term dynamic range of continuous speech. *Journal of Speech, Language, and Hearing Research* 51 (2): 471–484.
- Hickson, L., and D. Byrne. 1997. Consonant perception in quiet: Effect of increasing the consonant-vowel ratio with compression amplification. *Journal of the American Academy of Audiology* 8: 322–332.
- Holube, I., V. Hamacher, and M.C. Killion. 2016. Multi-channel compression: Concepts and (early but timeless) results. *Hearing Review* 23 (2): 20–26.
- Hornsby, B.W., and T.A. Ricketts. 2001. The effects of compression ratio, signal-to-noise ratio, and level on speech recognition in normal-hearing listeners. *Journal of the Acoustical Society of America* 109 (6): 2964–2973.
- Humes, L.E. 2002. Factors underlying the speech-recognition performance of elderly hearing-aid wearers. *Journal of the Acoustical Society of America* 112 (3): 1112–1132.
- Humes, L.E., and J.R. Dubno. 2010. Factors affecting speech understanding in older adults. In *The Aging Auditory System, Chap. 8*, ed. S. Gordon-Salant, R.D. Frisina, A.N. Popper, and R.R. Fay, 211–257. New York, NY: Springer.
- IEC 60268-13. 1985. *Sound System Equipment—Part 13: Listening Tests on Loudspeakers*. International Electrotechnical Commission.
- Jenstad, L.M., and P.E. Souza. 2005. Quantifying the effect of compression hearing aid release time on speech acoustics and intelligibility. *Journal of Speech Language, and Hearing Research* 48 (3): 651–667.
- Jenstad, L.M., and P.E. Souza. 2007. Temporal envelope changes of compression and speech rate: Combined effects on recognition for older adults. *Journal of Speech Language, and Hearing Research* 50 (5): 1123–1138.
- Jerlvall, L., and A. Lindblad. 1978. The influence of attack time and release time on speech intelligibility: A study of the effects of AGC on normal hearing and hearing impaired subjects. *Scandinavian Audiology Supplementum* 6: 341–353.
- Jørgensen, S., and T. Dau. 2011. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *Journal of the Acoustical Society of America* 130 (3): 1475–1487.
- Jørgensen, S., S.D. Ewert, and T. Dau. 2013. A multi-resolution envelope-power based model for speech intelligibility. *Journal of the Acoustical Society of America* 134 (1): 436–446.

- Kates, J.M. 1993. Optimal estimation of hearing-aid compression parameters. *Journal of the Acoustical Society of America* 94 (1): 1–12.
- Kates, J.M. 2005. Principles of digital dynamic-range compression. *Trends in Amplification* 9 (2): 45–76.
- Kates, J.M. 2010. Understanding compression: Modeling the effects of dynamic-range compression in hearing aids. *International Journal of Audiology* 49 (6): 395–409.
- Kates, J.M., and K.H. Arehart. 2014. The hearing-aid speech quality index (HASQI) version 2. *Journal of Audio and Engineering Society* 62 (3): 99–117.
- Keidser, G., H. Dillon, M. Flax, T. Ching, and S. Brewer. 2011. The NAL-NL2 prescription procedure. *Audiology Research* 1 (1)
- Keidser, G., K. Rohrseitz, H. Dillon, V. Hamacher, L. Carter, U. Rass, and E. Convery. 2006. The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers. *International Journal of Audiology* 45 (10): 563–579.
- Killion, M.C., H. Teder, A.C. Johnson, and S.P. Hanke. 1992. Variable recovery time circuit for use with wide dynamic range automatic gain control for hearing aid. US Patent 5,144,675.
- Kowalewski, B., J. Zaar, M. Fereczkowski, E.N. MacDonald, O. Strelcyk, T. May, and T. Dau. 2018. Effects of slow- and fast-acting compression on hearing-impaired listeners' consonant-vowel identification in interrupted noise. *Trends in Hearing* 22: 1–12.
- Kramer, S.J. 2008. *Audiology: Science to Practice*, 1st ed. Plural Publishing.
- Kreisman, B.M., A.G. Mazeveski, D.J. Schum, and R. Sockalingam. 2010. Improvements in speech understanding with wireless binaural broadband digital hearing instruments in adults with sensorineural hearing loss. *Trends in Hearing* 14 (1): 3–11.
- Kryter, K.D. 1962a. Methods for the calculation and use of the articulation index. *Journal of the Acoustical Society of America* 34 (11): 1689–1697.
- Kryter, K.D. 1962b. Validation of the articulation index. *Journal of the Acoustical Society of America* 34 (11): 1698–1702.
- Kuk, F.K. 1996. Theoretical and practical considerations in compression hearing aids. *Trends in Amplification* 1 (1): 5–39.
- Kuklasinski, A., S. Doclo, S.H. Jensen, and J. Jensen. 2016. Maximum likelihood PSD estimation for speech enhancement in reverberation and noise. *IEEE/ACM Transactions on Audio, Speech, Language Processing* 24 (9): 1599–1612.
- Lai, Y.-H., P.-C. Li, K.-S. Tsai, W.-C. Chu, and S.-T. Young. 2013. Measuring the long-term SNRs of static and adaptive compression amplification techniques for speech in noise. *Journal of the American Academy of Audiology* 24 (8): 671–683.
- Lopez-Poveda, E.A. 2014. Why do I hear but not understand? Stochastic undersampling as a model of degraded neural encoding of speech. *Frontiers in Neuroscience* 8: 348.
- Ma, N., T. May, and G.J. Brown. 2017. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, Language Processing* 25 (12): 2444–2453.
- May, T. 2018. Robust speech dereverberation with a neural network-based post-filter that exploits multi-conditional training of binaural cues. *IEEE/ACM Transactions on Audio, Speech, Language Processing* 26 (2): 406–414.
- May, T., B. Kowalewski, and T. Dau. 2018. Signal-to-noise-ratio aware dynamic range compression in hearing aids. *Trends in Hearing* 22: 1–12.
- May, T., B. Kowalewski, M. Fereczkowski, and E.N. MacDonald. 2017. Assessment of broadband SNR estimation for hearing aids. In *Proceedings of the ICASSP*, 231–235.
- May, T., S. van de Par, and A. Kohlrausch. 2013. Binaural localization and detection of speakers in complex acoustic scenes. In *The Technology of Binaural Listening, Chap. 15*, ed. J. Blauert, 397–425. Berlin, Germany: Springer.
- Middlebrooks, J.C., and D.M. Green. 1991. Sound localization by human listeners. *Annual Review of Psychology* 42 (1): 135–159.

- Moore, B.C., and B.R. Glasberg. 1988. A comparison of four methods of implementing automatic gain control (AGC) in hearing aids. *British Journal of Audiology* 22 (2): 93–104.
- Moore, B.C.J. 2008. The choice of compression speed in hearing aids: Theoretical and practical considerations and the role of individual differences. *Trends in Amplification* 12 (2): 103–112.
- Moore, B.C.J., R.W. Peters, and M.A. Stone. 1999. Benefits of linear amplification and multichannel compression for speech comprehension in backgrounds with spectral and temporal dips. *Journal of the Acoustical Society of America* 105 (1): 400–411.
- Moore, B.C.J., and A. Sek. 2013. Comparison of the CAM2 and NAL-NL2 hearing aid fitting methods. *Ear and Hearing* 34 (1): 83–95.
- Musa-Shufani, S., M. Walger, H. von Wedel, and H. Meister. 2006. Influence of dynamic compression on directional hearing in the horizontal plane. *Ear and Hearing* 27 (3): 279–285.
- Naylor, G., and R.B. Johannesson. 2009. Long-term signal-to-noise ratio at the input and output of amplitude-compression systems. *Journal of the American Academy of Audiology* 20 (3): 161–171.
- Neher, T., G. Grimm, V. Hohmann, and B. Kollmeier. 2014. Do hearing loss and cognitive function modulate benefit from different binaural noise-reduction settings? *Ear and Hearing* 35 (3): e52–e62.
- Neuman, A.C., M.H. Bakke, C. Mackersie, S. Hellman, and H. Levitt. 1998. The effect of compression ratio and release time on the categorical rating of sound quality. *Journal of the Acoustical Society of America* 103 (5): 2273–2281.
- Neumann, J. 2008. Method for dynamic determination of time constants, method for level detection, method for compressing an electric audio signal and hearing aid, wherein the method for compression is used. US Patent 7,333,623.
- Noble, W., D. Byrne, and K. Ter-Horst. 1997. Auditory localization, detection of spatial separateness, and speech hearing in noise by hearing impaired listeners. *Journal of the Acoustical Society of America* 102 (4): 2343–2352.
- Noble, W., K. Ter-Horst, and D. Byrne. 1995. Disabilities and handicaps associated with impaired auditory localization. *Journal of the American Academy of Audiology* 6: 129–129.
- Oxenham, A.J., and S.P. Bacon. 2004. Psychophysical manifestations of compression: Normal-hearing listeners. In *Compression: From Cochlea to Cochlear Implants, Chap. 3*, ed. S.P. Bacon, A.N. Popper, and R.R. Fay, 62–106. New York, NY: Springer.
- Pavlovic, C.V., and G.A. Studebaker. 1984. An evaluation of some assumptions underlying the articulation index. *Journal of the Acoustical Society of America* 75 (5): 1606–1612.
- Plomp, R. 1988. The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function. *Journal of the Acoustical Society of America* 83 (6): 2322–2327.
- Rana, B., and J.M. Buchholz. 2018. Effect of audibility on better-ear glimpsing as a function of frequency in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America* 143 (4): 2195–2206.
- Reinhart, P.N., P.E. Souza, N.K. Srinivasan, and F.J. Gallun. 2016. Effects of reverberation and compression on consonant identification in individuals with hearing impairment. *Ear and Hearing* 37 (2): 144–152.
- Rhebergen, K.S. 2006. Modeling the speech intelligibility in fluctuating noise. Ph.D. thesis, Faculty of Medicine, University of Amsterdam, Amsterdam, The Netherlands.
- Rhebergen, K.S., T.H. Maalderink, and W.A. Dreschler. 2017. Characterizing speech intelligibility in noise after wide dynamic range compression. *Ear and Hearing* 38 (2): 194–204.
- Rhebergen, K.S., N.J. Versfeld, and W.A. Dreschler. 2008. Quantifying and modeling the acoustic effects of compression on speech in noise. *Journal of the Acoustical Society of America* 123 (5): 3167–3167.
- Rhebergen, K.S., N.J. Versfeld, and W.A. Dreschler. 2009. The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise. *Journal of the Acoustical Society of America* 126 (6): 3236–3245.
- Robles, L., and M.A. Ruggero. 2001. Mechanics of the mammalian cochlea. *Physiological Reviews* 81 (3): 1305–1352.

- Rosen, S. 1992. Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London, Series B* 336 (1278): 367–373.
- Scollie, S., R. Seewald, L. Cornelisse, S. Moodie, M. Bagatto, D. Lournagaray, S. Beaulac, and J. Pumford. 2005. The desired sensation level multistage input/output algorithm. *Trends in Amplification* 9 (4): 159–197.
- Sherbecoe, R.L., and G.A. Studebaker. 2003. Audibility-index predictions of normal-hearing and hearing-impaired listeners' performance on the connected speech test. *Ear and Hearing* 24 (1): 71–88.
- Shi, L.-F., and K.A. Doherty. 2008. Subjective and objective effects of fast and slow compression on the perception of reverberant speech in listeners with hearing loss. *Journal of Speech, Language, and Hearing Research* 51 (5): 1328–1340.
- Simonsen, C., and T. Behrens. 2009. A new compression strategy based on a guided level estimator. *Hearing Review* 16 (13): 26–31.
- Sockalingam, R., M. Holmberg, K. Eneroth, and M. Shulte. 2009. Binaural hearing aid communication shown to improve sound quality and localization. *Hearing Journal* 62 (10): 46–47.
- Souza, P. 2016. Speech perception and hearing aids. In *Hearing Aids, Chap. 6*, ed. G.R. Popelka, B.C.J. Moore, R.R. Fay, and A.N. Popper, 151–180. New York: Springer.
- Souza, P., and F. Gallun. 2010. Amplification and consonant modulation spectra. *Ear and Hearing* 31 (2): 268–276.
- Souza, P., E. Hoover, and F. Gallun. 2012a. Application of the envelope difference index to spectrally sparse speech. *Journal of Speech Language, and Hearing Research* 55 (3): 824–837.
- Souza, P., R. Wright, and S. Bor. 2012b. Consequences of broad auditory filters for identification of multichannel-compressed vowels. *Journal of Speech Language, and Hearing Research* 55 (2): 474–486.
- Souza, P.E. 2002. Effects of compression on speech acoustics, intelligibility, and sound quality. *Trends in Amplification* 6 (4): 131–165.
- Souza, P.E., K.T. Boike, K. Witherell, and K. Tremblay. 2007. Prediction of speech recognition from audibility in older listeners with hearing loss: Effects of age, amplification, and background noise. *Journal of the American Academy of Audiology* 18 (1): 54–65.
- Souza, P.E., L.M. Jenstad, and K.T. Boike. 2006. Measuring the acoustic effects of compression amplification on speech in noise. *Journal of the Acoustical Society of America* 119 (1): 41–44.
- Souza, P.E., L.M. Jenstad, and R. Folino. 2005. Using multichannel wide-dynamic range compression in severely hearing-impaired listeners: Effects on speech recognition and quality. *Ear and Hearing* 26 (2): 120–131.
- Souza, P.E., and V. Kitch. 2001. The contribution of amplitude envelope cues to sentence identification in young and aged listeners. *Ear and Hearing* 22 (2): 112–119.
- Souza, P.E., and C.W. Turner. 1996. Effect of single-channel compression on temporal speech information. *Journal of Speech, Language, and Hearing Research* 39 (5): 901–911.
- Souza, P.E., and C.W. Turner. 1998. Multichannel compression, temporal cues, and audibility. *Journal of Speech, Language, and Hearing Research* 41 (2): 315–326.
- Souza, P.E., and C.W. Turner. 1999. Quantifying the contribution of audibility to recognition of compression-amplified speech. *Ear and Hearing* 20 (1): 12–20.
- Stelmachowicz, P.G., J. Kopun, A. Mace, D.E. Lewis, and S. Nittrouer. 1995. The perception of amplified speech by listeners with hearing loss: Acoustic correlates. *Journal of the Acoustical Society of America* 98 (3): 1388–1399.
- Stone, M.A., and B.C. Moore. 1999. Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses. *Ear and Hearing* 20 (3): 182–192.
- Stone, M.A., and B.C. Moore. 2002. Tolerable hearing aid delays. II. Estimation of limits imposed during speech production. *Ear and Hearing* 23 (4): 325–338.
- Stone, M.A., and B.C. Moore. 2003. Effect of the speed of a single-channel dynamic range compressor on intelligibility in a competing speech task. *Journal of the Acoustical Society of America* 114 (2): 1023–1034.

- Stone, M.A., and B.C. Moore. 2004. Side effects of fast-acting dynamic range compression that affect intelligibility in a competing speech task. *Journal of the Acoustical Society of America* 116 (4): 2311–2323.
- Stone, M.A., and B.C. Moore. 2007. Quantifying the effects of fast-acting compression on the envelope of speech. *Journal of the Acoustical Society of America* 121 (3): 1654–1664.
- Stone, M.A., and B.C. Moore. 2008. Effects of spectro-temporal modulation changes produced by multi-channel compression on intelligibility in a competing-speech task. *Journal of the Acoustical Society of America* 123 (2): 1063–1076.
- Stone, M.A., B.C. Moore, J.I. Alcántara, and B.R. Glasberg. 1999. Comparison of different forms of compression using wearable digital hearing aids. *Journal of the Acoustical Society of America* 106 (6): 3603–3619.
- Stone, M.A., and B.C.J. Moore. 1992. Syllabic compression: Effective compression ratios for signals modulated at different rates. *British Journal of Audiology* 26 (6): 351–361.
- Strelcyk, O., N. Nooraei, S. Kalluri, and B. Edwards. 2012. Restoration of loudness summation and differential loudness growth in hearing-impaired listeners. *Journal of the Acoustical Society of America* 132 (4): 2557–2568.
- van Buuren, R.A., J.M. Festen, and T. Houtgast. 1999. Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality. *Journal of the Acoustical Society of America* 105 (5): 2903–2913.
- Van Tasell, D.J. 1993. Hearing loss, speech, and hearing aids. *Journal of Speech, Language, and Hearing Research* 36 (2): 228–244.
- Varga, A.P., and H.J.M. Steeneken. 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication* 12 (3): 247–251.
- Verschuure, J., A. Maas, E. Stikvoort, R. De Jong, A. Goedegebure, and W. Dreschler. 1996. Compression and its effect on the speech signal. *Ear and Hearing* 17 (2): 162–175.
- Villchur, E. 1973. Signal processing to improve speech intelligibility in perceptive deafness. *Journal of the Acoustical Society of America* 53 (6): 1646–1657.
- Villchur, E. 1989. Comments on “the negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function” [J. Acoust. Soc. Am. 83, 2322–2327 (1988)]. *Journal of the Acoustical Society of America* 86 (1): 425–427.
- Walaszek, J. 2008. Effects of compression in hearing aids on the envelope of the speech signal, signal based measures of the side-effects of the compression and their relation to speech intelligibility. Master’s thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark.
- Wang, D.L., and J. Chen. 2018. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26 (10): 1702–1726.
- Wiggins, I.M., and B.U. Seeber. 2011. Dynamic-range compression affects the lateral position of sounds. *Journal of the Acoustical Society of America* 130 (6): 3939–3953.
- Wiggins, I.M., and B.U. Seeber. 2012. Effects of dynamic-range compression on the spatial attributes of sounds in normal-hearing listeners. *Ear and Hearing* 33 (3): 399–410.
- Wong, D.E., J. Hjortkjær, E. Ceolini, S.V. Nielsen, S.R. Grifol, S. Fuglsang, M. Chait, T. Lunner, T. Dau, S.-C. Liu, and A. de Cheveigné. 2018. A closed-loop platform for real-time attention control of simultaneous sound streams. In *ARO Midwinter Meeting (abstract)*.
- Woods, W.S., D.J. Van Tasell, M.E. Rickert, and T.D. Trine. 2006. SII and fit-to-target analysis of compression system performance as a function of number of compression channels. *International Journal of Audiology* 45 (11): 630–644.
- Yund, E.W., and K.M. Buckles. 1995a. Enhanced speech perception at low signal-to-noise ratios with multichannel compression hearing aids. *Journal of the Acoustical Society of America* 97 (2): 1224–1240.
- Yund, E.W., and K.M. Buckles. 1995b. Multichannel compression hearing aids: Effect of number of channels on speech discrimination in noise. *Journal of the Acoustical Society of America* 97 (2): 1206–1223.

- Yund, E.W., H.J. Simon, and R. Efron. 1987. Speech discrimination with an 8-channel compression hearing aid and conventional aids in background of speech-band noise. *Journal of Rehabilitation Research and Development* 24 (4): 161–180.
- Zaar, J., and T. Dau. 2016. Sources of variability in consonant perception and implications for speech perception modeling. In *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing* 437–446. Springer
- Zurek, P.M., and L. Delhorne. 1987. Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment. *Journal of the Acoustical Society of America* 82 (5): 1548–1559.

Index

A

- Absolute category rating (ACR), 403
- Absorption
- effective surface area, 259
 - total, 259, 278
- Abstraction, 206, 209, 213–215
- Accuracy, 64, 81, 83
- Acoustic τ , 640
- Acoustic reverberation, 462
- Acoustics, 397
- Acoustic scene, 33
- Active inference, 119
- Active listening, 365, 666, 677, 691
- Active noise cancellation or control (ANC), 356, 358
- Activity map binaural, 265
- Adaptation, 159, 203, 205, 207, 209, 214, 218
- binaural, 206
 - restart, 206
- Aesthetics
- awareness, 456
 - musical, 456
 - of recorded sounds, 455
- Affective cognition, 384, 666
- Afferent (ascending), 4
- Agent
- multimodal, 6
 - robotic, 98
- Age-related effect, 77, 79, 83
- Aibo, Sony, 383
- Akaike's criterion, 310
- Algorithmic music, 384
- AlloSphere, 353
- Ambisonics, 233, 360, 376, 384, 440, 581
- energy vector, 607
 - higher-order (HOA), 581, 585, 586, 588–590, 593–595, 599, 601, 608, 611–613
 - near-field-compensated higher-order (NFC-HOA), 581, 596, 608, 611
 - velocity vector, 607
- American Bureau of Standards, 438
- Amplification, 376, 378
- linear, 764
 - nonlinear, 764
- Amplitude-to-rate-code conversion, 10
- AmpMe, 383
- Anatomical transfer function (ATF), 364, 368, 371, 384
- Anechoic, 74, 80, 83, 739, 748, 749
- Anechoic orchestra recordings, 189
- Anyware, 380
- Apparent source width (ASW), 138, 612, 774, 790
- Applications, 654
- Arousal, 458
- Arousal-Valence plane, 469
- Arrangement, 397
- Artificial intelligence (AI), 382, 383
- Asimo, Honda, 383
- ASMR. *See* autonomous sensory meridian response (ASMR)
- Assistive technology, 356, 383, 667
- Association model, 600, 601
- Association module, 493
- Attend, 377–380
- Attention, 17, 409, 418, 419, 424, 570
- auditory, 7

- elicitation, 406
 - focus of, 493
 - focusing of, 18
 - Attentional selection, 54
 - Attention reorienting, 698–700, 708
 - Attention switch, 64, 67–69, 83
 - intentional , 73
 - paradigm, 65, 66
 - involuntary , 64
 - Attenuation, 374–376, 378, 385
 - Attribute elicitation, 407
 - Audio definition model (ADM), 385
 - Audio engineering, 458
 - Audiophile, 465
 - Audio steganography, 383
 - Audio windowing, 351, 352, 376, 381
 - Audiovisual localisation, 291
 - Audiovisual object, 699, 701, 710–714, 719, 720, 725
 - Auditory
 - front-end, 496
 - maps, 50
 - objects, 33
 - Auditory adaptation, 633, 646, 654
 - HRTF, 633
 - localization, 633
 - room acoustics, 633, 635, 642, 644
 - Auditory cortex (AC), 133, 161, 479, 700, 705
 - Auditory cues
 - dynamic, 637, 639
 - spectral, 639
 - Auditory display, 666
 - 3D, 667, 676, 692
 - dynamic, 692
 - Auditory effects, 629
 - Auditory illusion, 582, 584, 591, 600, 603, 609, 615
 - drone effect, 605
 - octave illusion, 603
 - scale illusion, 605
 - spatial (SAI), 582, 584, 606, 610, 613–615
 - Auditory information processing, 64
 - Auditory localization, 596, 607, 608, 611
 - Auditory modeling, 217
 - binaural, 212
 - Auditory nerve (AN), 129
 - Auditory object, 116
 - Auditory perception, 638, 639, 644, 654
 - Auditory periphery, 272
 - Auditory scene, 692, 745, 747, 748
 - analysis (ASA), 34, 92, 119, 161, 377, 408, 418, 491, 582, 600, 602, 605
 - analysis, computational (CASA), 34
 - analysis, computational (CASA), 492, 493
 - Auditory scene presentation, 455
 - Auditory scene reproduction
 - head-related, 458
 - room-related, 458
 - stereophonic, 459
 - surround, 459
 - Auditory selective attention, 62, 64, 71, 74, 81, 83
 - Auditory source width (ASW), 354, 355
 - Auditory stream, 116
 - segregation, 492
 - Audium, 353
 - Augmented acoustic reality, 655
 - Augmented reality (AR), 358, 361, 382, 654
 - Auralization, 175, 229, 381
 - Auro-3D, 385
 - Authentic reproduction, 624
 - Authenticity, 70, 71, 614
 - Autocorrelation, 268
 - Auto-correlogram analysis, 48
 - Autofocus, 379, 380
 - Automatic speech recognition (ASR), 512, 518
 - basic principles, 518
 - deep learning approaches, 521
 - feature extraction, 518
 - hidden Markov models, 519
 - impact of additive noise, 512
 - impact of reverberation, 512
 - technology, 512
 - Autonomous sensory meridian response (ASMR), 375
 - Average variance extracted (AVE), 445
 - Avoidance
 - behavior, 689
 - collision, 689
- B**
- Background noise, 398, 734, 741
 - Bassiness, 441
 - Bayesian
 - estimation, 325
 - inference theory, 326
 - method, 93
 - model, 326
 - network, 99
 - Bayesian causal inference, 301

- Bayesian inference, 38
 - Bayesian rule, 298
 - BBBeat, 678, 686
 - Beamforming, 352, 356, 735–742, 746
 - Beamforming algorithms, 525
 - delay-and-sum beamforming, 525
 - minimum-variance distortionless re-
sponse (MVDR), 525
 - spatial aliasing, 525
 - Behavior, 403–405, 424
 - Berlin philharmonic orchestra, 439
 - Better-ear
 - approach, 19
 - listening, 549
 - Big data, 383
 - Binaural, 62, 69, 152
 - listening paradigm, 81, 83
 - activity map, 267, 268, 279
 - manikin, 276
 - mixing console, 496
 - room impulse response, 257, 266, 271
 - Binaural cues, 737–742
 - Binaural de-reverberation, 551
 - Binaural display, 351, 356, 358, 359, 362–
365, 367, 371, 666, 692
 - class diagram, 677
 - dynamic, 676, 692
 - latency, 677
 - Binaural hearing
 - application to ASR, 513
 - Binaural incoherence, 124
 - Binaural intelligibility level difference
(BILD), 741
 - Binaural level, 184
 - Binaurally-integrated cross-correlation au-
tocorrelation mechanism (BICAM),
265, 268
 - Binaural modeling, 208
 - Binaural quality index (BQI), 138
 - Binaural reproduction
 - individual, 72
 - methods, 74
 - non-individual, 74
 - quality, 71
 - Binaural room impulse response (BRIR),
128, 232, 445, 462, 646
 - artificial, 630, 647
 - individual, 629, 635, 647
 - synthesis, 653
 - Binaural signal, 666
 - Binaural synthesis, 444, 646
 - auditory scene, 646
 - dynamic, via headphones, 623, 627, 641
 - room-related, 646
 - Binaural technology, 692
 - Binaural technology for robust ASR, 526
 - approaches based on deep learning, 536
 - complementarity of additive interference
and reverberation, 532
 - early approaches, 527
 - mask estimation based on EC processing,
535
 - mask estimation based on interaural co-
herence, 533
 - mask estimation based on ITD and IID,
530
 - mask estimation based on onset empha-
sis, 531
 - Binaural unmasking, 549
 - Binding, 499
 - hypotheses, 500
 - Birmingham electroacoustic sound theatre
(BEAST), 353
 - Blackbird Studio C, 264
 - Blackboard system, 7, 23, 92
 - acoustic-cues layer, 100
 - architecture, 94, 100
 - computational framework, 95
 - confusion-hypothesis layer, 102
 - for solving complex problems, 93
 - localization-hypothesis layer, 102
 - perceptual-hypothesis layer, 103
 - Blindness, 677
 - congenitally, 680
 - Blind source separation (BSS), 377
 - Bluetooth, 750
 - Blumlein, 585
 - Bokeh*, 356
 - Bone conduction, 352, 360
 - Bose Frames. *See* Frames, Bose
 - Bottom-up
 - information, 17
 - processing, 4
 - Bottom-up processing, 44
 - Boundary element method (BEM), 356
 - Boundary surface control (BoSC), 356, 667
 - Boundary-surface control. *See* BoSC
 - Brachium of the inferior colliculus, 132
 - Brainstem, 152
 - Brilliance, 445
 - British broadcasting corporation (BBC), 438
 - BS.1116, 403
- C**
- Cardboard, Google, 375

- Causal inference
 - after audiovisual stimulation, 308
 - with a generative model, 301
 - without a generative model, 304
- Cave acoustics, 254, 263, 264
- Cepstral mean normalization (CMN), 519
- Change deafness, 55
- Channel-based encoding, 384, 385
- Chirp, 383
- Clarity, 228, 231, 353
- Classical concert venues, 437
- Cleveland Orchestra, 237
- Clifton effect, 207, 213
- Club Transmediale, 353
- Cochlea, 764
- Cochlear implants (CI), 352, 358, 734, 750
- Cocktail party, 66, 71
 - effect, 18, 139, 624
 - situations, 55
- Code
 - spatial, 151
- Coding channels, 158
- Cognition, 415, 416
- Cognitive control mechanisms, 62, 65, 83
- Cognitive factors, 571
- Cognitive map, 381, 385, 642, 679
- Coherence, 739, 742–744
 - interaural, 260
 - temporal, 39
- Coherent-to-diffuse energy ratio (CDR), 534
- Coherent-to-diffuse ratio weighting (CDRW), 534
- Coincidence detection, 154
- Coloration, 413, 415, 425
- Combination models, 562
- Communication, 397, 404, 687
 - high-definition, 692
 - interpersonal, 687
- Comparative fit indices (CFIs), 445
- Completeness, 613
- Composer, 465
- Comprehensibility, 397
- Compression, 10
 - ratio (CR), 769
 - threshold (CT), 769
- Compress sensing, 119
- Computational auditory scene analysis (CASA), 121, 523
 - approaches using deep learning, 536
 - traditional approaches, 528, 531, 533, 535
- Computational SNR models, 554
- Concepts, 435
- Conceptual representation, 436
- Concert hall, 173
 - binaural dynamic responsiveness, 194
 - diffuse reflection, 185
 - direct sound, 177
 - dynamic responsiveness, 194
 - early reflections, 177, 182
 - lateral reflections, 181
 - late reverberation, 180, 197
 - rankings, 438
 - spatial responsiveness, 174
- Cone of confusion, 73, 126, 274, 319
- Confirmatory factor analyses (CFA), 445
- Congruence, 291, 701, 710, 712–714, 719, 722, 725
 - Congruence, perceptual, 21
- Congruency, 493
- Congruency effect, 67, 68, 79, 81
- Congruent, 67
- Consistence of stimulation, 296
- Construct reliability (CR), 445
- Context, 207, 214, 216, 398, 425
 - context-specific, 164
 - listening, 400, 403, 420
 - parameters, 624, 631, 646, 654
- Contextual modulation, 165
- Continuity illusion, 51
- Continuous quality scales (CQS), 471
- Controlled listening experiment, 462
- Convolution, 353
- Coordinate system
 - head-related, 275
 - room-related, 275
- Cophase and subtract, 17
- Corollary discharge, 335
- Correlation-based models, 560
- Correlation coefficient, 670
 - bi-dimensional, 682
- CR
 - effective (ECR), 781, 787
 - nominal, 781, 786, 788
- Critical band, 352
- Critical distance, 259
- Cross-channel correlation, 48
- Cross correlation, 251, 268, 397
 - interaural (ICC), 262, 272
- Cross-modal bias, 325
- Crossmodal cue, 351, 352, 354, 369, 381, 384, 385
- Crosstalk cancellation, 74, 357
- Crosstalk compensation. *See* crosstalk cancellation
- Crowdsourcing, 423

- Cue
 - auditory, 690
 - communication, 687
 - crossmodal, 666
 - dynamic, 670
 - mobility, 686
 - primitive grouping cues, 35
 - schema-based, 35
 - sound localization, 670
 - top-down, 51
 - visual, 677, 687, 689
- Cue-reliability hypothesis, 324
- Cue-stimulus interval, 68
- Cultural artifacts, 435
- Cyberphysical system, 382

- D**
- Data-over-sound, 383
- Deafen, 377–380
- Decay rate
 - exponential, 271
- Decision device, 273
- Deepfake, 384
- Deep learning, 383
 - for ASR, 521, 536
- Deep neural network (DNN), 16, 494, 497
- Definition D_{50} , 353
- Delay, 394
- Depth perception, 639
- Dereverberation, 735, 736, 743, 745
- Dichotic, 62, 65, 69
 - listening paradigm, 83
- Differences
 - interaural, 152
- Diffraction, 374
- Diffusion, 352, 354, 355, 374, 382
- Digital signal processing (DSP), 234, 677
- Diminished reality, 356, 377
- Dip listening, 548
- Directional audio coding (DirAC), 356, 360, 440
- Direction of arrival (DOA), 463
- Directivity coefficient, 260
- Direct methods, 403
- Direct-to-reverberant ratio (DRR), 534, 635, 641, 644, 646, 744, 750, 775
 - estimation, 780
 - reduction, 775
- Distance, 406
- Distance-based amplitude panning (DBAP), 356
- Distance perception, 639, 644

- Distractor, 64, 70
- Distributed mode actuator (DMA), 352
- Distributed mode loudspeaker (DML), 352
- Dolby Atmos, 385
- Dominance, 297
- Doppler effect, 677
- Dorsal cochlear nucleus (DCN), 130, 152
- DTS:X, 385
- Dual multiple factor analysis (DMFA), 241
- Dual-process model of aesthetic experience, 468
- Dual-resonance nonlinear model (DRNL), 13
- Dynamic
 - binaural display, 692
 - cross-talk cancellation (CTC), 233
 - dynamic-range, 438
 - dynamic-weighting model, 21
 - ear input, 691
 - time warping, 646
 - weighting, 493
- Dynamic-range compression, 351, 375

- E**
- Eardrum impedance
 - middle-ear impedance, 11
- Ear-filter bank, 14
- Ear plug, 72
- Eartop display, 360
- Echo, 353, 371, 376
 - threshold, 203, 204, 214, 626
- Echolocation, 219, 640
- Edutainment, 667, 677, 692
- Efferent, 4, 335
- Ego-sphere, 493
- Electroencephalography (EEG), 403, 608
- Elevation, 263, 268
- Emotion, 411
- Emotional response
 - annotation, 469
 - inhibition, 475
- Empathetic computing, 382, 383
- Empathic computing. *See* empathetic computing
- Endogenous, 68, 77
- Energetic masking, 548
- Energy decay curve (EDC), 653
- Energy-time curve (ETC), 122
- Engagement, 182
- Ensemble conditions, 227
- Envelope
 - distortion, 771

temporal, 771
 Envelopment, 198, 440, 475
 Environment
 reverberant, 204, 213, 215, 216
 Environment classification
 acoustic, 95
 Equalization, 461
 Equalization-cancellation, 17, 517, 549, 535
 Evolution, 155
 Exogenous, 68, 77
 Expectation, 396, 398, 420, 633, 648
 Experience, 396, 398, 404, 413, 633
 prior, 402
 Experimental aesthetics, 466
 Experimental Media and Performing Arts
 Center (EMPAC), 264
 Experimental paradigm, 63
 Expert, 397, 406
 Exploration, 637, 645
 Exploratory factor analysis (EFA), 445
 Exploratory movement, 5
 Extended reality (XR), 356, 358, 360–362,
 377, 382
 Externalization, 626, 630, 633, 635, 637,
 646, 649, 774, 785
 Eye contact, 687, 692
 Eye tracker, 751–753

F

F0 segregation, 551
 Face contact, 687
 Factor analysis, 438
 Features, 405, 407, 418, 735–737, 741, 745–
 747, 749
 auditory, 406
 schema-based, 51
 Feedback, 423, 634
 cognitive, 23
 loops, 4, 161
 mechanism, 98
 reflective, 21
 reflexive, 4
 sensory-motor, 19
 Feelings, 396, 419
 Field studies, 225
 Filterbank
 gammatone, 785
 STFT, 767
 Filter theory, 63
 Finite difference time domain (FDTD), 232
 Finite elements method (FEM), 232
 Finite impulse-response (FIR) filter, 258

Fission

boundary, 39
 obligatory, 40
 streaming, 39
 Fixed-mobile convergence (FMC), 383, 384
 Flash profiling (FP), 471
 Fletcher, 585
 Floor reflection, 228, 233
 Flow, 383
 Fluency
 conceptual, 482
 perceptual, 482
 Fluency assessor, 482
 Focused source, 589, 596, 599
 Focus group, 444
 Frame-of-reference hypothesis, 324
 Frames, Bose, 358
 Frame-theory, 437
 Free choice profiling (FCP), 471
 Free-energy principle, 118
 Frontal (coronal) plane, 372
 Front/back confusion, 16, 20, 218, 275, 279,
 497
 Front-back reversal, 339, 357, 366, 367
 Functional magnetic resonance imaging
 (fMRI), 605, 608
 Functional model, 456
 Functional near-infrared spectroscopy
 (fNIRS), 608
 Fundamental frequency (F0), 550
 Fusion, 204, 208, 213
 obligatory, 40
 streaming, 39

G

Gain
 calculation, 769
 make-up, 769, 786
 prescription rule, 769
 Game
 action, 678, 686, 692
 auditory, 679, 686
 maze, 678, 679, 692
 racing, 678
 Gammatone
 filter, 14
 filterbank, 268
 Gammatone-frequency cepstral coefficients
 (GFCC), 536
 Gaussian-mixture models, 50
 General slowing, 78, 80
 Geometrical acoustics (GA), 232
 Geometric model, 256, 277
 Geotagging, 374

- Gestalt, 7, 33
 - perception, 37
 - rules, 33, 492
 - theory, 120
- Gestalt rules
 - closure, 37
 - common fate, 37
 - proximity, 37
 - similarity, 37
- Glimpse proportion, 565
- Glimpses, 549
- Glimpsing, 42, 548
 - models, 564
 - spectral, 551
- Global positioning system (GPS), 376, 384, 752
- Google Cardboard. *See* Cardboard, Google
- Google Tone, 383
- Graphical models, 24
- Ground truth, 474, 498
- Grouping, 162
- Groupware, 382
- Guide dog, 685
- Gulbenkian Grande Auditorio, 227

- H**
- Harmonic cancellation, 551
- Haydn Saal, 256, 264
- Hazard score, 502
- Head
 - cues, 335
 - head-tracking, 358, 376
 - motion, 403, 413, 418
 - turning, 666, 667
- Head-and-torso simulator (HATS), 7, 100
- Head-mounted displays (HMD), 358, 360, 362, 382, 383, 676
- Head movements, 206, 218, 263, 272, 330, 666, 717–725, 727
- Headphone
 - equalization, 71
 - extra-aural, 647
 - transfer function (HPTF), 71, 630, 647
- Head-related
 - impulse responses (HRIRs), 123, 496, 666, 677
 - individual, 71, 83
 - non-individual, 73, 77, 83, 634
 - transfer function (HRTF), 71, 122, 191, 258, 319, 364, 368, 371, 496, 514, 601, 633, 666, 676, 739, 742, 748, 749, 751
- Head rotation, 273, 274, 366–369, 372, 404, 410, 413, 666
 - active, 674
 - angle, 275
 - anti-phase, 668
 - in-phase, 668
- Head-turning modulation (HTM), 21
- Hearable, 356, 360
- Hearing, 33
 - impaired-listeners, 33
 - impairment, 733–735, 740, 741, 744, 750, 754
 - instruments, 733
 - loss, 733, 764, 786
- Hearing aid, 356, 360, 382, 764
 - behind the ear (BTE), 734
 - completely in the canal (CIC), 734
 - in the canal (ITC), 734
 - in the ear (ITE), 734, 740
 - programs, 737, 746, 752, 753
 - receiver in the ear canal (RIC), 734
 - wireless, 735, 736, 739, 740, 745
- Hearsay-II system, 99
- Heart rate, 403, 753
- Hedonic response, 478
- Helmholtz, 583
- Helmholtz reciprocity, 380
- Hemispheric population coding, 159
- Heuristics
 - rule-based, 93
- Hidden Markov model (HMM), 53, 519
- High-density loudspeaker array (HDLA), 353
- High-order ambisonics (HOA), 667
- Hohle-Fels cave, 254, 264, 268
- Holophony, 585, 586
 - higher-order ambisonics (HOA), 459
 - wave-field synthesis (WFS), 459
- Homo
 - neandertalis, 253
 - sapiens, 252
- Honda Asimo. *See* Asimo, Honda
- Horizontal plane, 70, 372
- Hypothesis driven, 5

- I**
- IC histogram, 785
- Identification, 397, 419, 423
- Ill-posed problem, 676, 692
- Illusion
 - auditory, 623, 625, 655
- Image
 - splitting, 205
- Image split, 774, 785, 789
- IMAX, 360

- Immersion, 397, 475, 623
 - Importance, 710, 711, 714, 720, 725, 727
 - Impulse response, 397
 - Incongruence, perceptual, 21
 - Incongruency, 67, 493
 - Indirect methods, 403
 - Individualization, 631
 - Individual vocabulary profiling (IVP), 471
 - Inertial measurement unit (IMU), 369, 375
 - Inferior colliculus (IC), 131, 152
 - Inferior frontal cortex (IFC), 133
 - Inferior parietal lobe (IPL), 133
 - Information
 - auditory, 666, 678, 689
 - nonverbal, 687
 - visual, 689
 - Informational masking, 548
 - In-head-localization (IHL), 626, 636
 - Inhibition, 155
 - of irrelevant information, 63, 67, 77, 79, 83
 - of return dynamic, 493
 - Inhibitory deficit theory, 78
 - Initial time-delay gap (ITDG), 375, 646
 - modification, 653
 - Input-gain control, 10
 - Input/output function, 769, 788
 - Inside-the-head localization or locatedness (IHL), 364
 - In-situ measurements, 228
 - Integration, 698–701, 703
 - Integration window, 292
 - Intelligibility, 394, 397, 423
 - Intelligibility, short-time objective, 560
 - Interaural cross-correlation (IACC), 353
 - Interaction
 - active, 643
 - authentic, 643, 645
 - non-authentic, 643, 645
 - passive, 643
 - Interaural
 - arrival-time difference (ITD), 100
 - coherence (IC), 774
 - level difference (ILD), 75, 100, 774
 - time difference, 75
 - Interaural coherence, 124, 556
 - Interaural cross-correlation coefficient (IACC), 124
 - Interaural cross-correlation function, 124
 - Interaural intensity difference (IID), 354, 371
 - Interaural level difference (ILD), 124, 155, 318, 549, 737, 739, 742–745, 748, 749, 751
 - distortion, 774
 - histogram, 785
 - Interaural phase differences (IPDs), 126
 - Interaural-polar coordinate system, 121
 - Interaural time difference (ITD), 124, 155, 263, 267, 318, 354, 371, 549, 737–739, 742–745, 748, 749
 - Interferer, periodic, 569
 - Internal model, 118
 - Internal reference, 463
 - Internet of things (IoT), 361, 382, 383, 385
 - Interpolation, 653
 - Interquartile distance (IQD), 649
 - Intimacy, 182, 445
 - Inverse effectiveness, 291
 - Inverse model of recent experience, 308
 - Isophones
 - equal-loudness contours, 10
 - ITU audio definition model (ADM). *See* audio definition model (ADM)
- J**
- Just noticeable difference (JND), 136, 641, 649
- K**
- Kansei* engineering, 384
 - Knowledge source (KS), 92, 94, 95, 481, 495
- L**
- Labeled hedonic scale (LHS), 471
 - Laboratory experiments, 225
 - Latency, 668, 677
 - Latent measurement models, 443
 - Lateral angle, 333
 - Lateral-energy fraction, 397
 - Lateral superior olive (LSO), 131, 153
 - Layer model, 397
 - Learning
 - reinforcement, 494
 - reinforcement, 96
 - supervised, 95
 - Learning processes, 627
 - Least squares fitting, 309
 - Level estimation, 767
 - Linguistic,
 - encoding, 435
 - relativism, 437

- Listener, 399, 411
 Listener envelopment (LEV), 138, 612
 Listening
 active, 666, 677, 691
 Listening test, 394, 401, 404
 Listening zone. *See* sound zone
 Liveliness, 445
 Localization, 73, 75, 397, 406, 413, 425, 635, 637, 733, 735, 737, 741, 744, 745, 747–749, 753
 auditory, 15, 497
 dominance, 204
 error, 631, 633
 in azimuth, 291
 in distance, 294
 in elevation, 293
 in reverberation, 204
 Location-based service (LBS), 361
 Lombard effect, 373
 Loudness, 397, 406, 415, 437, 440, 442, 445
 Loudspeaker orchestra, 175
 Loudspeaker response, 480
 Lyric Speaker, 381
- M**
- Machine hearing, 93
 Machine learning. *See* artificial intelligence (AI)
 Machine listening, 33
 Machine translation, 383
 Magnetoencephalography (MEG), 608
 Mammals, 151
 Map, neuronal, 151
 Markov
 hidden models, 99
 random fields, 99
 Mask
 binary, 36
 ideal binaural (IBM), 41
 ideal-ratio mask (IRM), 43
 Masking
 energetic, 52
 informational, 52
 Master quality authenticated (MQA), 381
 Mastering, 461
 Matched filter, 47
 Maximum likelihood estimation, 299
 McGurk effect, 624
 Meaning
 allocation, 499
 assignment of, 397, 418
 Measurement invariance, 443
 Medial geniculate body (MGB), 133
 Medial-olivocochlear reflex (MOCR), 12
 Medial superior olive (MSO), 131, 153
 Median plane, 73, 80
 Median (sagittal) plane, 368, 372
 Mel frequency cepstral coefficients (MFCC), 519
 MEMS. *See* microelectromechanical system (MEMS)
 Mentalmapper, 678, 679
 Merge AR/VR Headset, 375
 Micro-electrical-mechanical-systems (MEMS) microphones, 750
 Microelectromechanical system (MEMS), 375
 Microsoft Soundscape, 374
 Middle-ear-muscle reflex (MEMR), 11
 Minimum audible angle (MAA), 135, 645, 675
 Minimum-audible-movement angle (MAMA), 645
 Minimum-audible-movement distance (MAMD), 646
 Missing data, 712, 713, 715–717
 Mixed reality (MR), 351, 358, 377
 Mo-cap (motion capture), 381, 382
 Mobile-ambient interface, 352, 360, 361, 384
 Modality appropriateness, 298
 Modality precision, 298
 Mode
 emulation, 500
 idle, 498
 patrol, 501
 Modelling
 sound localisation, 297
 sound localisation after audiovisual stimulation, 307
 sound localisation during audiovisual stimulation, 297
 Models
 of binaural interaction, 514
 cross-correlation-based models, 515
 equalization-cancellation (EC) model, 517
 Gaussian-mixture, 102
 graphical, 99
 Lindemann model, 517
 position-variable model, 517
 stereausis model, 517
 that assume cue integration, 297
 that do not assume cue integration, 301
 Modulation-based models, 558

- Modulation spectrum, 558, 782, 787
- Monaural parameters, 229
- Motion
 - capture. *See* mo-cap (motion capture)
 - motion-parallax effect, 640
 - motion-tracked binaural (MTB), 647
 - parallax, 342
 - sensor, 674
- Motivation, 711, 712, 714, 721
- Motor reactions, 699–701, 703, 705, 706, 708, 709, 711, 714
- Moving
 - listeners, 330
 - sound sources, 330
- Moving minimum-audible angle, 340
- MPEG-H, 385
- Multiactuator panel (MAP), 352
- Multi-channel Wiener filter (MWF), 738–740
- Multidimensional, 405–407, 418, 421
 - scaling (MDS), 436, 438, 470
- Multilayer perceptron (MLP), 521
- Multimodal, 699–703, 708, 709, 711–714, 727
 - interaction, 351, 352, 354, 365, 369, 375, 381, 384, 385
- Multimodal-fusion-&-inference model, 21
- Multiple stimuli with hidden reference and anchor (MUSHRA), 471, 612
- Multipresence, 351, 352, 379, 380
- Multisensory
 - integration, 292
 - perception, 591
- Multi-talker situation, 68
- MUSHRA, 403, 408, 409
- Music
 - dynamics, 186
 - spectrum, 189
- Musician, 465
- Music information retrieval (MIR), 236
- Music performance analysis, 236
- Music-source separation, 53
- Mute, 377–380
- N**
- Narrowcasting, 351, 352, 377, 379, 380, 384
- Natural language processing (NLP), 384
- Naturalness of auralization methods, 242
- Nearest loudspeaker synthesis, 233
- Nearphone, 360
- Neural-motor control, 337
- Neural substracts, 291
- Neurophysiology, 466
- New Adventures in Sound Art, 353
- No interaction, 301
- Noise reduction, 736, 738–740, 742, 743, 745, 746
- Non-auditory modality, 463
- Non-diegetic music, 370
- Nonsonorealistic rendering, 369
- Non-verbal information, 371
- Null-steering antenna, 17, 18
- O**
- Object
 - assigning meaning to an, 23
 - auditory, 15
 - auditory identification, 499
 - auditory localization, 497
 - formation, 15, 34, 497, 500
 - hypotheses, 501
 - object-based representation, 54
 - perceptual, 15, 497
 - selection, 34
- Object-based encoding, 371, 376, 384, 385
- Object formation, 397, 418
- Obstruction, 353, 374
- Occlusion, 280, 353, 374
- Oculus Quest, 375
- Olivary complex, 11
- Omnidirectional sound source, 259, 260, 278
- On-/offset analysis, 49
- Open dome, 71
- Open guided sound (OGS) earphones, 359
- Open science, 413
- Oriented-gradients detector, 505
- Otoacoustic emission (OAE), 358
- Overhead reflector, 227
- Overlap-add (OLA), 769
- P**
- Paired comparisons, 409, 436
- Panasonic Wear Space. *See* Wear Space, Panasonic
- Panoramic potentiometer (Pan-pot), 356, 381
- Pantophonic speaker array, 353, 355
- Pepper, SoftBank, 383
- Percept, 436
- Perception, 151, 352, 353, 362, 364, 365, 367, 369
 - dual nature of, 404
- Perceptionism, 365
- Perceptual

assessment, 585, 592–594, 606
 inference, 118
 linear prediction (PLP), 519
 mixing time, 653
 model, 635
 Performance, 403, 405, 412, 413
 cost, 66, 68, 76, 83
 venues, 436
 Peripheral autonomic nervous system (PANS), 608
 Periphonic speaker array, 353
 Personalized auditory reality (PARTy), 655
 Personal sound amplification product (PSAP), 356, 360, 382
 Perspective, 397
 Phantom source, 328, 582, 586, 607
 Phantom-walker illusion, 639
 Phase-difference channel weighting algorithm (PDCW), 532
 Phase inversion technique, 780
 Phonemic restoration, 51
 Physical modeling, 384
 Pinna factor, 337
 Pipe organ, 239
 Pitch, 397
 Plane
 horizontal, 255
 Plasticity, 151
 context-dependent, 151
 Plausibility, 70, 71, 458, 614, 623, 626, 633, 645
 Point of operation, 10, 13
 Point of subjective straight ahead (PSSA), 673
 Polar coordinate system, 319
 Position, 397, 398, 420
 Positional disparity, 642
 Position detector
 visual, 506
 Power-normalized cepstral coefficients (PNCC), 519
 Prägnanz, 37
 Precedence effect, 10, 185, 203–205, 263, 265, 381, 596, 604, 626, 635
 break-down, 212
 build-up, 207, 211
 dynamic, 208
 Prediction, 699, 704, 705, 708
 error, 552
 Predictive coding, 118
 Preference mapping, 405, 407, 415
 Premotor cortex (PMC), 133
 Presence, 614

Primary auditory cortex (A1), 133
 Principal-components analysis (PCA), 610
 Procedural audio, 384
 Process
 dynamic, 208
 Projection mapping, 361, 385
 Proprioception, 218, 337, 382
 Proto-event, 23
 Proximity, 182, 441
 Pseudo inverse matrix, 310
 Pseudophones, 295
 Pseudophony, 357, 365, 366
 Public address (PA), 360

Q

Quadraphony, 581, 586
 Quality
 assumed, 398
 basic audio, 408, 409, 411
 content, 397
 dialogue, 397
 perceived, 633
 product-sound, 397
 sound, 393, 395, 426
 Quality and aesthetic judgment, 456
 Quality-of-experience (QoE), 393, 396, 400, 426, 594, 609
 Quality-of-service (QoS), 395
 Quantified self, 384

R

Radiation, 373, 378
 Rate map, 15
 Ray tracing, 251, 255, 256
 signal flow, 258
 software, 256
 Reaction time, 64, 81, 83
 Realism, 458
 Real-time auralization, 233
 Recorded-reproduced sound, 456
 Rectification
 halfwave, 273
 Reference, 397–400, 405
 internal, 633, 648
 Reflection, 353, 371, 374, 375
 coefficient, 269
 diffuse, 251, 258, 264
 specular, 264
 Reflective feedback, 4, 21
 Reflective processes, 120
 Reflexive feedback, 4
 Reflexive processes, 120

Refraction, 374
 Reinforcement of notes, 437
 Relative spectral analysis (RASTA), 519
 Remapping, 273
 Remote microphone, 735, 736, 739, 745, 748–751
 Repertory-grid technique (RGT), 610
 Representation, 151
 Reproducible research, 413, 425
 Reproduction
 binaural, 637
 via headphones, 635
 via loudspeaker, 627
 Response-cue interval, 67
 Reverberance, 227, 438, 440, 442, 445, 449
 Reverberant tail, 260, 268, 278
 Reverberation, 353, 359, 371, 374–376, 406, 415, 418, 437, 512, 551
 time (RT), 80, 81, 83, 231, 254, 258, 397, 414, 438
 Reverberation time RT_{60} , 353
 Reverb time RT_{60} . *See* reverberation time RT_{60}
 Reverse-hierarchy theory, 21
 RoBoHon, Sharp, 383
 Robot, 362, 383, 699, 708–710, 712–717, 719, 721, 727
 Robust speech recognition, 518
 Room-acoustical predictors, 241
 Room acoustical quality inventory (RAQI), 444, 448
 Room acoustic impression, 443
 Room acoustic parameters
 adjustment, 646
 clarity, 641
 energy-based, 646, 652
 measurement, 646
 reverberation time, 641
 simulation, 646
 time-based, 646
 Room acoustics, 409, 414, 415
 Room adaptation, 569
 Room coloration, 552
 Room-divergence effect (RDE), 625, 631, 635, 644, 646–648, 652
 Room learning, 203, 206, 209, 214, 215, 218
 Room transfer function (RFT), 676
 Root-mean-square errors of approximation (RMSEA), 445
 Roughness, 397
 Royal Festival Hall, London, 438

S
 Saccadic suppression, 676, 692
 Sagittal plane, 333
 Saliency, 21, 701–704, 706–708, 710, 721
 map, 493
 semantic, 21
 Scattering, 374
 Scene, 408, 415, 418, 423
 acoustic, 400, 414
 auditory, 7
 aural, 397
 Scene-based encoding, 384, 385
 Scheduler, 94, 495
 Schemata, primitive, 7
 Sea of Sound, 353
 Segmentation, 418, 420, 421
 Segregation, 33, 159
 obligatory, 40
 pitch-based, 49
 primitive stream, 40
 Selection criterion, 67
 Self-assessment-manikin (SAM), 469
 Self-motion, 218, 637
 visual effects, 638
 Self rotation, 336
 Semantic differential, 437
 Semantic scales, 436
 Sensor
 accelerometer, 751–754
 ballistocardiogram, 753
 blood pressure, 753
 electroencephalogram (EEG), 752–754
 electrooculography (EOG), 751
 GPS, 752, 753
 gyroscope, 751
 heart rate, 753
 temperature, 753
 Sensory, 416
 assessor, 482
 compensation, 677
 descriptors, 470
 evaluation, 402, 405, 410
 memory, 419
 weights, 305
 weights with causal inference, 305
 Shadow filtering, 780
 Shadowing task, 63
 Sharpening the ears, 17
 Sharpness, 373, 397
 Sharp RoBoHon. *See* RoBoHon, Sharp
 Short-time discrete Fourier transform (STFT)
 analysis, 766

- synthesis, 769
- SiFASo, 667
- Signal
 - direct, 261
 - direct-reverberant ratio, 461
 - driven, 5
 - dynamic range, 462
 - running, 263, 265, 268
 - spectral balance, 462
 - timbre, 462
- Signal processing for robust ASR, 522
 - computational auditory scene analysis (CASA), 523
 - homomorphic deconvolution, 522
 - missing feature approaches, 523
 - spectral subtraction, 522
 - vector Taylor series (VTS), 522
- Signal-to-noise ratio (SNR), 512, 548
 - broadband, 773, 786
 - estimation, 777
 - in the envelope power domain, 784, 788
 - reduction, 773
- Signal-to-spike rate conversion, 10
- Similarity judgments, 436
- Simultaneous localization and mapping (SLAM), 369
- Simultaneous organization, 50
- Situation awareness, 373, 381, 383
- Skill
 - communication, 687, 688
 - sound localization, 686, 688, 689
- Skin conductance, 403
- Smart speaker, 351, 381–385
- SNR-based model, 553, 556
- SoftBank Pepper. *See* Pepper, SoftBank
- Solo (select), 378–380
- Sony Aibo. *See* Aibo, Sony
- Sound
 - mixing, 399, 400, 405, 410
 - pressure, 394, 397, 405
- Sound/audio quality
 - assessment, 456
- Sound externalization, 137
- Soundfield synthesis (SFS), 356
- SoundFormular, 678
- Sound localization, 135, 152, 203, 214, 215, 291, 667, 686, 688, 689
 - accuracy, 691
 - after audiovisual stimulation, 295
 - front-back error, 669
 - horizontal plane, 667, 672
 - instantaneous, 676
 - just noticeable difference, 673
 - median-plane, 667
 - model of, 342
 - performance, 691
 - training, 688, 689
- Sound location, 151
- Sound qualities, 437
- Sound reinforcement (SR), 360, 370, 382
- Sound rendering, 666
- Soundscape, 351–353, 358, 360, 362, 363, 369, 376, 377, 379, 381, 382, 384, 385
- Sound source
 - direction, 631
 - directivity, 641
 - level, 646
 - moving, 645
 - rotation, 641
 - stationary, 639, 641, 645
- Sound-source localization. *See* auditory-object localization
- Sound strength, 228
- Sound zone, 382, 384
- Source-image localization, 462
- Space
 - auditory, 151
- Spaciousness, 138, 183, 397
- Spatial aliasing, 586, 590, 593
- Spatial audio, 393, 400, 408, 413, 581, 582, 584, 591, 599, 605–607, 609, 610, 612, 613, 615
- Spatial cues, 738, 739, 750, 751
- Spatial decomposition method (SDM), 176, 232
- Spatial hearing, 183, 191, 666, 691
- Spatial impression, 138
- Spatial location, 67
- Spatially oriented format for acoustics (SOFA), 123
- Spatial maps
 - multisensory, 342
- Spatial matching, 291
- Spatial perception, 666, 676, 677
- Spatial release from masking (SRM), 139, 549, 771
- Spatial resolution, 160
- Spatial room impulse response (SRIR), 229
- Spatial separation, 550
- Spatial transparency, 445
- Spatial unmasking, 139, 605
- Spatial updating, 332, 341
- Spatiotemporal information, 231
- Spatiotopically organized, 326
- Speaker array, 351–353, 384

- Spectral-shape cues, 319
- Speech, 404, 423
 - audibility, 770
 - automatic (ASR), 512
 - comprehension, 571
 - context, 216
 - formant transition, 216
 - in reverberation, 203, 216
 - intelligibility, 445, 548, 764
 - material, 568
 - reception threshold (SRT), 741, 742
 - recognition (SR), 383
 - speech-intelligibility index, 553
 - speech-transmission index, 558
- Spike rates, 158
- Stage acoustics, 228
- Standardized root-mean-square residual (SRMR), 445
- Stapedius, 11
- Stereophony, 581, 582, 584, 585, 595, 600, 607, 612
- Stereotelephony, 374
- Stimulation pattern, 652
- Stream, 34
 - perception, 34
- Streaming, 737, 738, 748, 749
- Strength, 440, 442
- Subjective impression, 227
- Subjective straight ahead, 672
 - point of, 673
- Summing localization, 204, 596, 599, 600
- Superior colliculus, 132, 291
- Superior olivary complex (SOC), 131
- Superior temporal gyrus (STG), 133
- Support parameters, 228
- Suppression of slowly-varying components and the falling edge of the power envelope (SSF), 532
- Surprise
 - auditory, 493
 - Bayesian, 493
- Surround sound, 360, 371, 581, 586
- Sweet spot, 585, 599
- Switch cost, 67, 76, 77
- System latency. *See* latency
- TeleHead, 667
- Telepresence, 353, 380, 383
- Telexistence. *See* telepresence
- Temporal, 398, 419
 - matching, 291
 - smearing, 551
- Tensor tympani, 11
- Test-retest reliability, 447
- Text-to-speech (TTS), 383
- Thalamus, 133
- Theory of processing fluency, 467
- 360 Reality Audio, Sony, 358
- Timbral adjustments, 238
- Timbral performance attributes, 238
- Timbre, 397, 407, 442
- Time constant
 - ANSI, 767, 788
 - attack, 767, 786
 - nominal, 767, 786
 - release, 767, 775, 786
- Time-energy information, 231
- Time-to-contact, 639
- Tonal balance, 397
- Tone, 437
- Tonotopically organized, 327
- Top-down
 - information, 17
 - processing, 4
- Torus of confusion, 126
- Training, 634, 652, 654
 - cognitive map, 679
 - sound localization, 686, 688, 689
- Transfer effect, 686, 692
- Transfer learning, 423
- Transfer of learning. *See* transfer effect
- Translation
 - head, 637, 641
 - listener, 637, 641
 - positional changes, 639
- Transmission channel, 460
- Transmission function, 397
- Transparency, 397
- Trapezoid body (TB), 131
- Tune Mob, 383
- Turn-to-reflex, 16, 493, 698
- Two-channel stereo mix, 460

T

- Target, 67, 70, 83
- Task-cuing method, 65
- Task switching, 65
- Taxonomic organization, 436
- Technologies, 458

U

- Ubicomp (ubiquitous computing), 361, 362, 382
- Ubiquitous computing. *See* ubicomp
- Ultrasonic, 357, 383

- Uncanny valley, 594
 - Uncertainty, 570
 - Usability, 397
 - Useful-to-detrimental ratio, 562
- V**
- Validity, 443
 - Variable
 - continuous-valued, 24
 - discrete-valued, 24
 - Vection, 638
 - Vector base amplitude panning (VBAP), 233, 356, 360, 611
 - Vector Taylor series (VTS), 522
 - Ventral cochlear nucleus (VCN), 130
 - Ventriloquism
 - aftereffect, 296
 - effect, 297, 322, 323, 366, 624
 - Vestibular
 - cues, 218
 - system, 336, 362, 367, 368
 - Virtual
 - acoustic scene, 633, 637, 645
 - environments, 139
 - performance studio (VPS), 234
 - Virtual reality (VR), 358, 360, 361, 375, 409, 412, 667
 - auditory, 678
 - game, 678
 - generator, 26
 - Virtual Singing Studio, 234
 - Vision, 699–703, 706, 717, 718, 726
 - Visual
 - bias, 325
 - capture, 297, 322
 - dominance hypothesis, 324
 - vs. auditory space, 328
 - Visual cortex, 700, 706
 - Visual cue, 70
 - Visual effects, 628, 629, 631
 - Visually impaired, 359, 667, 677, 679, 687, 692
 - Vocabulary profiling, 436
 - Voice-chat, 384
- W**
- Walkable virtual loudspeaker setup, 655
 - Walk-through, 255
 - forest, 260, 261, 272
 - office, 251, 272
 - Wallach azimuth illusion, 330
 - Wallach vertical illusion, 332
 - Wave-field synthesis (WFS), 233, 356, 360, 581, 585, 586, 588–591, 594–596, 599, 601, 608, 610–613
 - optimized phantom source imaging, 612
 - Way-finding, 354, 374
 - Wearable, 754
 - computing, 360, 362, 382
 - Wear Space, Panasonic, 358
 - Welfare system, 677
 - Wheel of concert hall acoustics, 441
 - Wide dynamic range compression
 - binaural, 774, 790
 - conventional, 766, 787
 - DRR-aware, 779, 790
 - SNR-aware, 777, 787
 - Window of cue compatibility, 300
 - World knowledge, 19
 - World model, 24
 - internal, 98
- X**
- XR (extended reality). *See* extended reality (XR)