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 evaluation[1] 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—

[1] The 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

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_14

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 decoloration (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

[3]The 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).

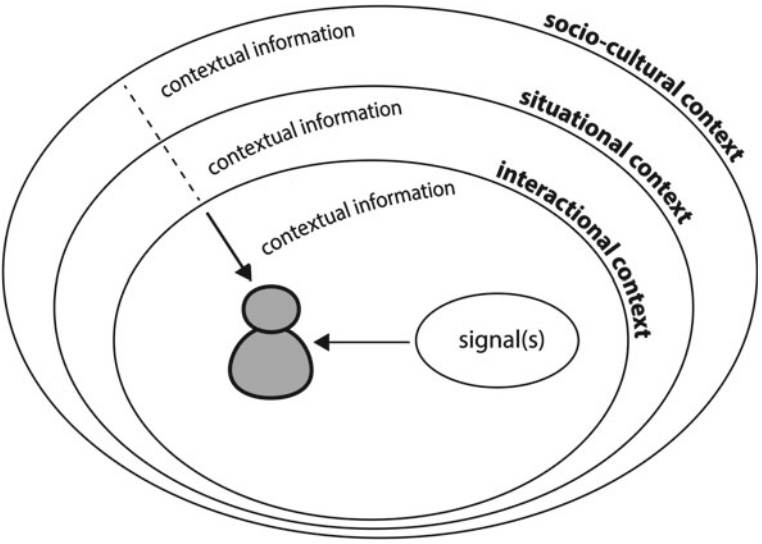

Fig. 1 Contexts of use of an audio-related system (Raake and Egger 2014), adopted from ideas by Geerts et al. (2010) and Moor (2012). The context-dependent roles of the persons, different implications of the physical environment, and of the other actors present in the different types of contexts determine their perception, as well as their evaluation of *Sound Quality* and *Quality of Experience*

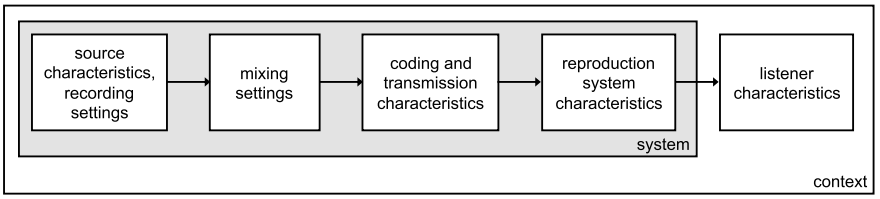

Fig. 2 End-to-end chain ("system") including sound recording, processing, transmission and reproduction in terms of the factors that ultimately determine *Sound Quality* and *Quality of Experience*—adapted from Wierstorf et al. (2018)

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 references[4] 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

[4]The representations available in memory in abstracted form, and used at the different perceptual and evaluation stages.

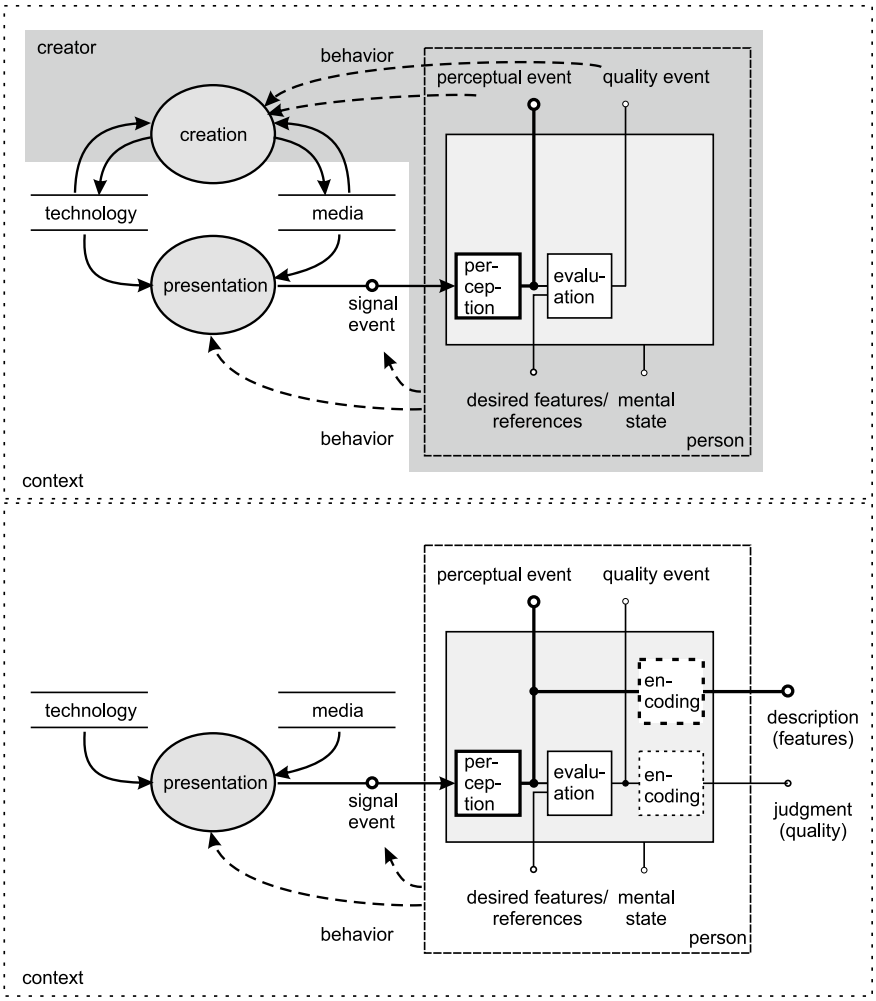

Fig. 3 **Top** Quality perception in the context of creation/production—adapted from Raake and Egger (2014). Perception and evaluation are represented in a simplified manner as two processing components in the mind of a person. The person and creator may be identical and *Quality of Experience* is used 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

[5] Often referred to as *subjective*, a somewhat misleading term avoided here.

[6] Often referred to as *objective*, erroneously implying that instrumental measurements bear objectivity, which they only do in case that they can be generalized.

[7] Note 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 Schoenenberg (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 (Schoenenberg 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: *Distance*	Binaural features used in near-field for distance perception (Blauert 1997; Zahorik et al. 2005)
Internality, externalization	Different acoustic, auditory and multimodal effects that determine the amount to which an auditory event is localized either *out-of-head* or *inside-the-head* (Hartmann and Wittenberg 1996; Blauert 1997; Brandenburg et al. 2020)
Localizability	Lateral/horizontal-plane localization (Blauert 1997). Related to *spatial fidelity* and respective modeling approaches as discussed in Rumsey et al. (2008) and Wierstorf et al. (2017a)
Depth, width, envelopment	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*.

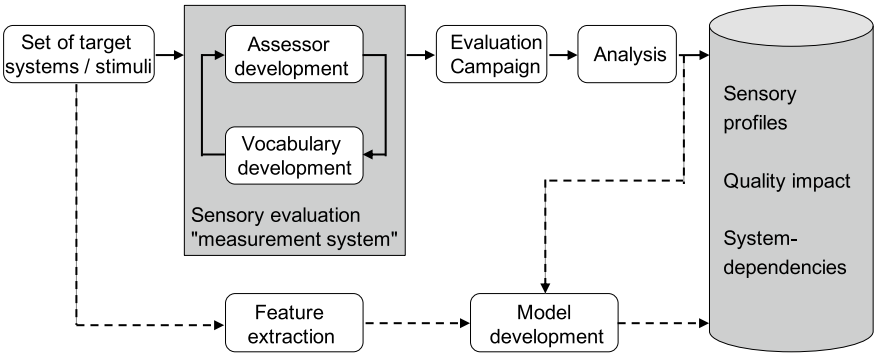

Fig. 4 Steps involved in the development of a sensory-evaluation-test method and, subsequently, of a sound quality model based on multidimensional analysis

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,

[8] Note 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

[9] It 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 Brüggen (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

[10] www.twoears.eu [last accessed, August 30, 2019].

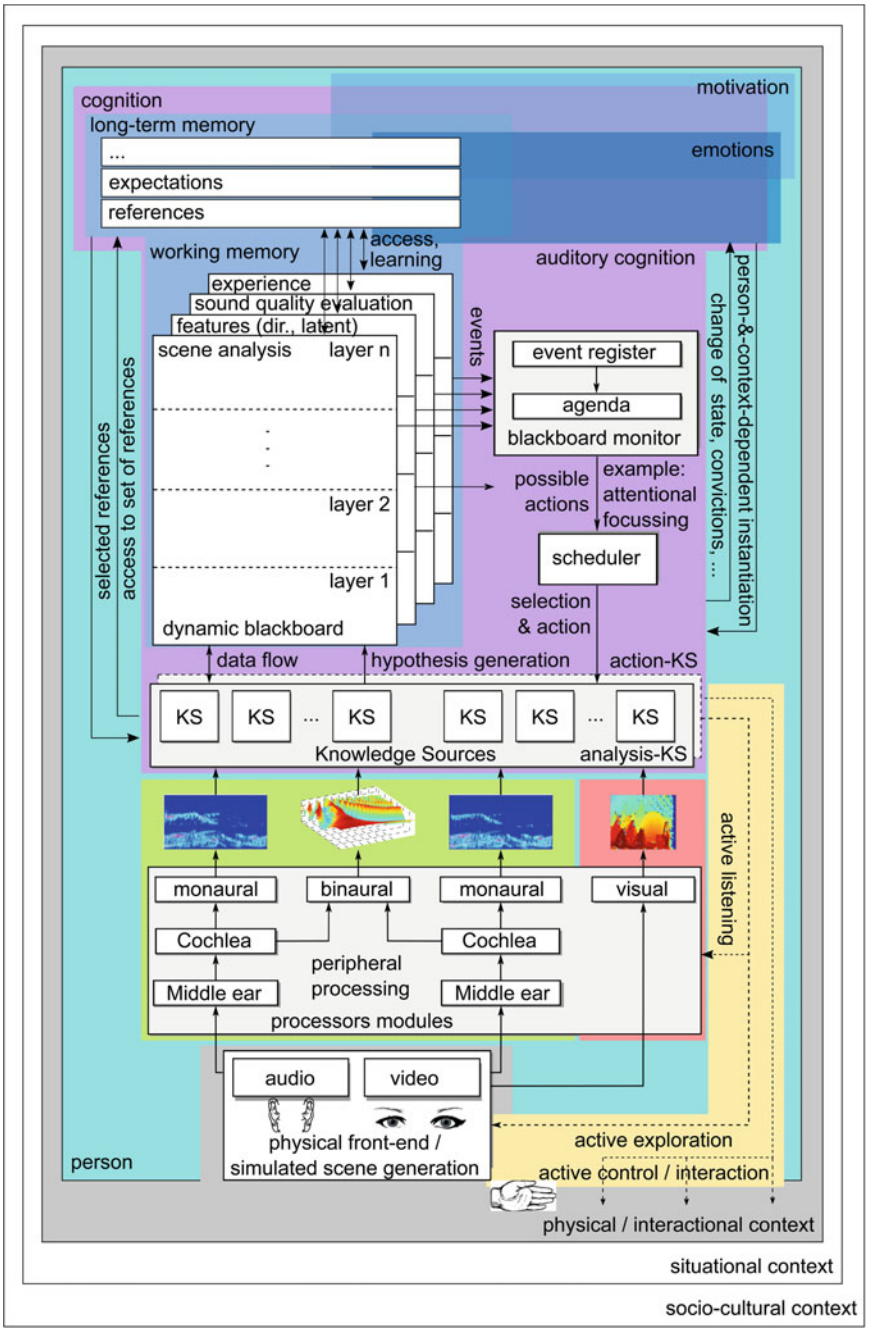

Fig. 5 Architecture of a comprehensive model of auditory and multimodal perception, *Sound Quality* and *Quality of Experience* formation. The picture is based on a conceptual drawing of a specifically tailored blackboard system (Raake and Blauert 2013), later amended by G. J. Brown and N. Ma in the 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, 3 D 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.

[11] Incorporating 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 Frond 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.

[12] See 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-sourcing[13] (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

[13]Crowd-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!EARS[15] 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.

Göring, 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 Vocabular*, 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 mediatized 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. Habels. 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. Meares. 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. Koeppe. 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 qoe 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 assessement 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.