Springer Handbook of Auditory Research

Kai Siedenburg
Charalampos Saitis
Stephen McAdams
Arthur N. Popper
Richard R. Fay *Editors*

Timbre: Acoustics, Perception, and Cognition

ASA PRESS

EXTRAS ONLINE

Springer

Springer Handbook of Auditory Research

Volume 69

More information about this series at http://www.springer.com/series/2506

The ASA Press

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

ASA PRESS

Kai Siedenburg • Charalampos Saitis
Stephen McAdams • Arthur N. Popper
Richard R. Fay
Editors

Timbre: Acoustics, Perception, and Cognition

ASA PRESS

Springer

Editors
Kai Siedenburg
Department of Medical Physics
and Acoustics
Carl von Ossietzky Universität Oldenburg
Oldenburg, Germany

Stephen McAdams
Schulich School of Music
McGill University
Montreal, QC, Canada

Richard R. Fay
Department of Psychology
Loyola University Chicago
Chicago, IL, USA

Charalampos Saitis
Audio Communication Group
Technische Universität Berlin
Berlin, Germany

Arthur N. Popper
Department of Biology
University of Maryland
Collage Park, MD, USA

ISSN 0947-2657 ISSN 2197-1897 (electronic)
Springer Handbook of Auditory Research
ISBN 978-3-030-14831-7 ISBN 978-3-030-14832-4 (eBook)
https://doi.org/10.1007/978-3-030-14832-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

The editors dedicate this volume to two giants in contemporary timbre research and practice: the late David L. Wessel (left), whose work on timbre by means of sound analysis and synthesis inspired some of the first computer music software that emphasizes real-time musical control of timbre, and the late Roger A. Kendall (right), whose explorations of perception, language, and meaning in musical timbre continue to resonate.

Acoustical Society of America

The purpose of the Acoustical Society of America (www.acousticalsociety.org) is to generate, disseminate, and promote the knowledge of acoustics. The Acoustical Society of America (ASA) is recognized as the world's premier international scientific society in acoustics, and counts among its more than 7000 members, professionals in the fields of bioacoustics, engineering, architecture, speech, music, oceanography, signal processing, sound and vibration, and noise control.

Since its first meeting in 1929, the ASA has enjoyed a healthy growth in membership and in stature. The present membership of approximately 7000 includes leaders in acoustics in the United States of America and around the world. The ASA has attracted members from various fields related to sound including engineering, physics, oceanography, life sciences, noise and noise control, architectural acoustics; psychological and physiological acoustics; applied acoustics; music and musical instruments; speech communication; ultrasonics, radiation, and scattering; mechanical vibrations and shock; underwater sound; aeroacoustics; macrosonics; acoustical signal processing; bioacoustics; and many more topics.

To assure adequate attention to these separate fields and to new ones that may develop, the Society establishes technical committees and technical groups charged with keeping abreast of developments and needs of the membership in their specialized fields. This diversity and the opportunity it provides for interchange of knowledge and points of view has become one of the strengths of the Society.

The ASA's publishing program has historically included *The Journal of the Acoustical Society of America, JASA-Express Letters, Proceedings of Meetings on Acoustics*, the magazine *Acoustics Today*, and various books authored by its members across the many topical areas of acoustics. In addition, ASA members are involved in the development of acoustical standards concerned with terminology, measurement procedures, and criteria for determining the effects of noise and vibration.

Series Preface

Springer Handbook of Auditory Research

The following preface is the one that we published in volume 1 of the Springer Handbook of Auditory Research back in 1992. As anyone reading the original preface, or the many users of the series, will note, we have far exceeded our original expectation of eight volumes. Indeed, with books published to date and those in the pipeline, we are now set for over 80 volumes in SHAR, and we are still open to new and exciting ideas for additional books.

We are very proud that there seems to be consensus, at least among our friends and colleagues, that SHAR has become an important and influential part of the auditory literature. While we have worked hard to develop and maintain the quality and value of SHAR, the real value of the books is very much because of the numerous authors who have given their time to write outstanding chapters and our many coeditors who have provided the intellectual leadership to the individual volumes. We have worked with a remarkable and wonderful group of people, many of whom have become great personal friends of both of us. We also continue to work with a spectacular group of editors at Springer. Indeed, several of our past editors have moved on in the publishing world to become senior executives. To our delight, this includes the current president of Springer US, Dr. William Curtis.

But the truth is that the series would and could not be possible without the support of our families, and we want to take this opportunity to dedicate all of the SHAR books, past and future, to them. Our wives, Catherine Fay and Helen Popper, and our children, Michelle Popper Levit, Melissa Popper Levinsohn, Christian Fay, and Amanda Fay Sierra, have been immensely patient as we developed and worked on this series. We thank them and state, without doubt, that this series could not have happened without them. We also dedicate the future of SHAR to our next generation of (potential) auditory researchers—our grandchildren—Ethan and Sophie Levinsohn, Emma Levit, and Nathaniel, Evan, and Stella Fay, and Sebatian Sierra-Fay.

Preface 1992

The Springer Handbook of Auditory Research presents a series of comprehensive and synthetic reviews of the fundamental topics in modern auditory research. The volumes are aimed at all individuals with interests in hearing research including advanced graduate students, postdoctoral researchers, and clinical investigators. The volumes are intended to introduce new investigators to important aspects of hearing science and to help established investigators to better understand the fundamental theories and data in fields of hearing that they may not normally follow closely.

Each volume presents a particular topic comprehensively, and each serves as a synthetic overview and guide to the literature. As such, the chapters present neither exhaustive data reviews nor original research that has not yet appeared in peer-reviewed journals. The volumes focus on topics that have developed a solid data and conceptual foundation rather than on those for which a literature is only beginning to develop. New research areas will be covered on a timely basis in the series as they begin to mature.

Each volume in the series consists of a few substantial chapters on a particular topic. In some cases, the topics will be ones of traditional interest for which there is a substantial body of data and theory, such as auditory neuroanatomy (Vol. 1) and neurophysiology (Vol. 2). Other volumes in the series deal with topics that have begun to mature more recently, such as development, plasticity, and computational models of neural processing. In many cases, the series editors are joined by a co-editor having special expertise in the topic of the volume.

Richard R. Fay, Chicago, IL, USA
Arthur N. Popper, College Park, MD, USA

SHAR logo by Mark B. Weinberg, Potomac, Maryland, used with permission

Volume Preface

Timbre is a foundational aspect of hearing. Roughly defined, timbre is thought of as any property other than pitch, duration, and loudness that allows two sounds to be distinguished. The remarkable ability of humans to recognize sound sources and events (e.g., glass breaking, a friend's voice, a tone from a piano) stems primarily from a capacity to perceive and process differences in the timbre of sounds. Timbre raises many important issues in psychology and the cognitive sciences, musical acoustics, speech processing, medical engineering, and artificial intelligence. Bringing together leading experts from around the world, this volume provides a joint forum for novel insights and the first comprehensive modern account of research topics and methods on the perception, cognition, and acoustic modeling of timbre.

This volume is the first dedicated to a comprehensive and authoritative presentation of the state of the art in research on timbre. Chapter 1, by the senior editors of this volume, gives an overview of the field, including a discussion of the various definitions of what timbre is. The chapter also gives a comprehensive overview of the book.

Following this, the next five chapters address the principal processes underlying timbre perception and cognition. In Chap. 2, Stephen McAdams discusses dimensional models of timbre based on multidimensional scaling (MDS) of timbre dissimilarity ratings and psychophysical explanations in terms of acoustical correlates of perceptual dimensions. Then, in Chap. 3, Trevor R. Agus, Clara Suied, and Daniel Pressnitzer describe the many important and intriguing empirical findings in the last 10 years on the categorization and recognition of sounds. In Chap. 4, Kai Siedenburg and Daniel Müllensiefen discuss research on long- and short-term memory for timbre. Chapter 5 by Charalampos Saitis and Stefan Weinzierl considers verbal descriptions of timbre and the rich semantic associations found in them. Following this, in Chap. 6, Vinoo Alluri and Sudarsana Reddy Kadiri review recent findings regarding the neural basis of timbre information processing from studies that used both animal models and human brain imaging.

The second part of this volume addresses specific scenarios of timbre perception. Chapter 7, by Samuel Robert Mathias and Katharina von Kriegstein, outlines important topics in voice processing and voice identification. Then, in Chap. 8, Stephen McAdams describes the various ways in which timbre shapes the perceptual experience of music. Following this, Jeremy Marozeau and Wiebke Lamping outline timbre perception in patients with severe or profound hearing loss who have received a cochlear implant (CI) in Chap. 9. Chapter 10 by Guillaume Lemaitre and Patrick Susini then focuses on the role of timbre in the evaluation of product sounds as related to the question of how sounds contribute to the aesthetic, functional, and emotional aspect of a product.

The third part of this volume is focused on the acoustic modeling of timbre. Chapter 11 by Marcelo Caetano, Charalampos Saitis, and Kai Siedenburg describes computational approaches to the acoustic description of sounds that have developed in the fields of psychoacoustics and music information retrieval to date. Then, in Chap. 12, Mounya Elhilali summarizes recent advances in the study and application of spectrotemporal modulation representations in speech and music. In the final chapter, Sølvi Ystad, Mitsuko Aramaki, and Richard Kronland-Martinet (Chap. 13) introduce an analysis-synthesis framework that derives intuitive control parameters of electronic sound synthesis directly from the statistics of input sounds.

Many of the chapters in this volume build on material in earlier volumes in the *Springer Handbook of Auditory Research*. Most notably, this first comprehensive treatment on the various aspects of timbre perception may serve as a natural complement to the *Springer Handbook of Auditory Research* volumes on the basic auditory parameters found in *Pitch: Neural Coding and Perception* (Volume 24, edited by Plack, Oxenham, Popper, and Fay, 2005) and *Loudness* (Volume 34, edited by Florentine, Popper, and Fay, 2011).

Kai Siedenburg, Oldenburg, Germany
Charalampos Saitis, Berlin, Germany
Stephen McAdams, Montréal, QC, Canada
Arthur N. Popper, College Park, MD, USA
Richard R. Fay, Chicago, IL, USA

Contents

Part I Principal Perceptual Processes

Part II Specific Scenarios

Part III Acoustical Modeling

Contributors

Trevor R. Agus School of Arts, English and Languages, Queen's University Belfast, Belfast, UK

Vinoo Alluri International Institute of Information Technology, Gachibowli, Hyderabad, India

Mitsuko Aramaki CNRS, Aix Marseille University, PRISM (Perception, Representations, Image, Sound, Music), Marseille, France

Marcelo Caetano Sound and Music Computing Group, INESC TEC, Porto, Portugal

Mounya Elhilali Laboratory for Computational Audio Perception, Center for Speech and Language Processing, Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA

Sudarsana Reddy Kadiri International Institute of Information Technology, Gachibowli, Hyderabad, India

Richard Kronland-Martinet CNRS, Aix Marseille University, PRISM (Perception, Representations, Image, Sound, Music), Marseille, France

Wiebke Lamping Hearing Systems Group, Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

Guillaume Lemaitre STMS-IRCAM-CNRS-UPMC, Paris, France

Jeremy Marozeau Hearing Systems Group, Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

Samuel Robert Mathias Neurocognition, Neurocomputation and Neurogenetics Division, Yale University School of Medicine, New Haven, CT, USA

Stephen McAdams Schulich School of Music, McGill University, Montreal, QC, Canada

Daniel Müllensiefen Department of Psychology, Goldsmiths, University of London, London, United Kingdom

Daniel Pressnitzer Laboratoire des Systèmes Perceptifs, Département d'études Cognitives, Paris Science & Lettres – PSL University, Centre national de la recherche scientifique, Paris, France

Charalampos Saitis Audio Communication Group, Technische Universität Berlin, Berlin, Germany

Kai Siedenburg Department of Medical Physics and Acoustics, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

Clara Suied Département de Neurosciences et Sciences Cognitives, Institut de recherche biomédicale des armées, Brétigny-sur-Orge, France

Patrick Susini STMS-IRCAM-CNRS-UPMC, Paris, France

Katharina von Kriegstein Technische Universität Dresden, Dresden, Germany

Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Stefan Weinzierl Audio Communication Group, Technische Universität Berlin, Berlin, Germany

Sølvi Ystad CNRS, Aix Marseille University, PRISM (Perception, Representations, Image, Sound, Music), Marseille, France

Chapter 1
The Present, Past, and Future of Timbre Research

Kai Siedenburg, Charalampos Saitis, and Stephen McAdams

Abstract Timbre is a foundational aspect of hearing. The remarkable ability of humans to recognize sound sources and events (e.g., glass breaking, a friend's voice, a tone from a piano) stems primarily from a capacity to perceive and process differences in the timbre of sounds. Roughly defined, timbre is thought of as any property other than pitch, duration, and loudness that allows two sounds to be distinguished. Current research unfolds along three main fronts: (1) principal perceptual and cognitive processes; (2) the role of timbre in human voice perception, perception through cochlear implants, music perception, sound quality, and sound design; and (3) computational acoustic modeling. Along these three scientific fronts, significant breakthroughs have been achieved during the decade prior to the production of this volume. Bringing together leading experts from around the world, this volume provides a joint forum for novel insights and the first comprehensive modern account of research topics and methods on the perception, cognition, and acoustic modeling of timbre. This chapter provides background information and a roadmap for the volume.

Keywords Acoustics · Auditory perception · History of auditory research · Music perception · Voice perception

K. Siedenburg (✉)
Department of Medical Physics and Acoustics, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
e-mail: kai.siedenburg@uni-oldenburg.de

C. Saitis
Audio Communication Group, Technische Universität Berlin, Berlin, Germany
e-mail: charalampos.saitis@campus.tu-berlin.de

S. McAdams
Schulich School of Music, McGill University, Montreal, QC, Canada
e-mail: stephen.mcadams@mcgill.ca

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_1

1.1 Timbre As a Research Field

The study of timbre has recently become the subject of a remarkable momentum. Much of this interest in timbre seems to emerge from several distinct research perspectives. First, psychophysical research into timbre has built novel pathways to investigating elementary questions regarding timbre's perceptual status. To what extent does timbre interact with pitch and loudness, and what role does it play in sound source recognition?

Second, cognitive neuroscience has increasingly addressed the psychophysical and neural bases of voice perception. What are the neural mechanisms and networks underlying the perception of arguably the most important auditory stimulus for humans?

Third, the field of music information retrieval has demonstrated new approaches to automatic musical-instrument recognition and genre classification from a biocognitive viewpoint. What are efficient computational representations of timbre that best mimic physiology and cognition?

Fourth, the research community is witnessing a strong musicological and music-theoretical interest in timbre. What are the conceptions and experiential dimensions of timbre that are shared between different periods and musical styles? What role does timbre play in nonclassical contexts, such as electroacoustic or popular music?

By probing those and related questions, numerous important and inspiring studies on timbre have been published in the decade prior to the writing of this overview. Moreover, no less than four independent workshops on timbre were organized between 2014 and 2018, reflecting the demand for direct discussions and exchange. The first small workshop in 2014 occurred at Telecom ParisTech (https://musictimbre.wp.imt.fr) with a focus on music information retrieval applications. This was followed by a meeting at Harvard University in 2015, the focus of which was on musicological issues. The Berlin Interdisciplinary Workshop on Timbre in 2017 at the Federal Institute for Music Research (*Staatliches Institut für Musikforschung*, http://www.timbre2017.tu-berlin.de) first brought together researchers from the diverse fields of science and humanities, specifically musicology, music cognition, cognitive neuroscience, and music information retrieval. This workshop gave rise to the idea of the present volume and most of its authors were part of the Berlin lineup. The scope was further expanded with perspectives from fields such as music composition, ethnomusicology, and sound recording at the conference "Timbre 2018: Timbre Is a Many-Splendored Thing" at McGill University in Montreal (https://www.mcgill.ca/timbre2018/), which received more than 130 paper submissions and was the largest conference on the topic so far. Reflecting aspects of this development, the upcoming *The Oxford Handbook of Timbre*, edited by Emily Dolan and Alexander Rehding (https://bit.ly/2PXgbQA) features historical, music-theoretical, and musicological perspectives on timbre.

This volume channels the momentum with regard to questions on perceptual and cognitive processing and acoustic modeling of timbre. As a result, it constitutes the first comprehensive treatment on the various aspects of timbre perception and will serve as a natural complement to the Springer Handbook of Auditory Research volumes on the basic auditory parameters of pitch (Plack et al. 2005) and loudness (Florentine et al. 2011).

1.1.1 Inter-Disciplinary Perspectives

Technically, timbre is a basic auditory attribute and should be of interest to all auditory scientists who are working on psychoacoustics, sound source perception, speech communication, soundscapes, or music. Given individual research traditions and foci, it is nonetheless unavoidable that the notion of timbre is encountered more frequently in some domains than in others, and individual research interests naturally bring about individualized perspectives.

Timbre permeates music listening, and polyphonic music often features aesthetically rich and intriguing treasures of timbre. In fact, the notion of timbre has a long-standing tradition in music perception research. In the nineteenth century, Helmholtz's (1877) seminal work outlined a theory of timbre that was dedicated to explaining the perception of musical-instrument sounds. Helmholtz used a simplifying short-hand definition that has become something akin to the textbook definition (with all its pitfalls, see Sect. 1.1.2): "By the quality of a tone [timbre, *Klangfarbe*] we mean that peculiarity which distinguishes the musical tone of a violin from that of a flute or that of a clarinet or that of the human voice, when all these instruments produce the same note at the same pitch" (Helmholtz 1877, p. 10). Perhaps for these reasons, much research framed under the headline of timbre has a particular eye on music perception (even though timbre has long been the neglected ugly duckling of music theory and musicology).

In speech, timbre plays a dual role. First, different speakers can be differentiated via timbre cues. Moreover, the sequences of phonemes that constitute speech beyond speaker information are based on timbral contrasts. Vowels differ by spectral envelope shape; consonants differ by spectrotemporal morphology. In other words, most of the meaning conveyed by speech is indeed transmitted via timbral contrast (although pitch also plays an essential role in tone languages). From this perspective, speech is a highly sophisticated system of timbral sequencing. Perhaps because this perspective is too general to be useful beyond speaker identity, one rarely observes connections being drawn in the literature between the vast field of speech research and basic psychoacoustic studies framed as timbre research (although see Patel 2008).

At the same time, timbre research, perhaps more than many other aspects of audition, relies on the integration of methods across fields. Helmholtz constitutes a prime example: he applied Fourier theory to the perception of acoustic signals and thereby integrated the state of the art in physics and auditory physiology. As will be further outlined in Sect. 1.2, progress in understanding timbre has not only been driven by smart and simple experiments, but also by advances in statistics (e.g., multidimensional scaling), signal processing (e.g., nonstationary signal analysis techniques such as the Short-Time Fourier Transform), or neurophysiology (e.g., brain imaging).

The chapters of this volume take inherently interdisciplinary perspectives but also reflect individual conceptual and methodological approaches toward timbre. Many examples stem from musical scenarios, but there are also dedicated discussions of general sound source recognition, voice perception and speaker identification, perception of industrial product sound quality, and timbre perception by cochlear implant users. Regardless of the specific application, the perceptual and cognitive processes addressed are of general significance.

1.1.2 Defining a Complex Auditory Parameter

A commonality at the heart of timbre research could be the willingness to focus on the direct and concrete sensory experience of sound while not considering the latter primarily as a medium to an otherwise abstract message in the form of strings of symbols, whether constituted via musical notation or linguistic categories. In the words of the musicologist Emily Dolan (2013):

> [Timbre] is the concept to which we must turn to describe the immediacies of how sounds strike our ears, how they affect us. It is the word we need when we want to discuss sound in terms of its particularities and peculiarities. To put it another way, to talk about timbre is to value sound as sound, and not as a sonic manifestation of abstract principles (Dolan 2013, p. 87).

Ironically, there may be another idea about timbre that auditory researchers agree on: that the concept is hard to define (cf., Krumhansl 1989; Siedenburg and McAdams 2017a). Perhaps for a lack of a better alternative, the American National Standards Institute (ANSI) definition is frequently revisited. For the sake of completeness (and tradition!):

> Timbre. That attribute of auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented and having the same loudness and pitch, are dissimilar [sic]. NOTE-Timbre depends primarily upon the frequency spectrum, although it also depends upon the sound pressure and the temporal characteristics of the sound (ANSI 1960/1994, p. 35).

Bregman (1990) severely criticized this definition, yet without providing any constructive alternative:

> This is, of course, no definition at all. […] The problem with timbre is that it is the name for an ill-defined wastebasket category. […] I think the definition … should be this: 'We do not know how to define timbre, but it is not loudness and it is not pitch.' […] What we need is a better vocabulary concerning timbre (Bregman 1990, pp. 92–93).

Comments such as these left many researchers in doubt as to whether the term is useful at all. In order to clear up some of the confusion around the notion of timbre, Siedenburg and McAdams (2017a) proposed four conceptual distinctions for the term. Here, these distinctions and potential implications are briefly outlined.

- *Timbre is a perceptual attribute.* It should be kept in mind that timbre is a perceptual attribute, as are pitch and loudness. Thus, it is only of limited use to speak of timbral properties of, say, an audio signal or musical orchestration without referring to the auditory sensation. In short, timbre lives not in the audio signal or in a musical score but in the mind of the listener.
- *Timbre is both a quality and a contributor to source identity.* This dual nature is often mentioned, but only rarely are the consequences of these subtleties considered. Regarding the qualitative stance, two sounds can be declared qualitatively dissimilar without bearing semantic associations or without their source/cause mechanisms being identified. On the other hand, timbre is defined as a collection of auditory sensory features that contributes to the inference (or specification) of sound sources and events. Importantly, timbral differences do not always correspond to differences in sound sources: Indeed, a single sound-producing object can give rise to a universe of timbres.
- *Timbre functions on different scales of detail.* There are differences in the granularity of timbral information: whereas the timbral differences between a bassoon played with different articulations may be subtle (or think of differences between a Stradivarius violin and a competitor model), the timbral differences between a bassoon and a piano are huge. Each of these separate timbral granularities or scales of detail encompasse interesting research questions.
- *Timbre is a property of fused auditory events.* Studies have begun to explore the acoustic correlates of what has been called "polyphonic timbre" (Alluri and Toiviainen 2010), defined as the *global sound* of a piece of music. In music information retrieval, it is common practice to run audio analyses on musical mixtures (also because automatic source separation is such a difficult computational problem). However, auditory scene-analysis principles should not be forgotten in this context. In fact, timbre may be viewed as a perceptual property of perceptually fused auditory events; if two or more auditory events do not fuse, they simply do not contribute to the same timbre. The simultaneously produced sounds from a bass drum, a handclap, and a synthesizer pad usually do not fuse into a single auditory image; as such, each of these sounds possesses an individual timbre in the mind of a listener. It is the emergent property of the combination of the individual timbres that evokes hip-hop, but there is no unitary "hip-hop timbre."

Whereas the first and last distinctions sharpen the notion of timbre, the second and third distinctions essentially acknowledge timbre as an umbrella term. The skeptical reader may insist that umbrella terms are too broad to be part of a refined scientific vocabulary. One might counter that there are other psychological concepts that are exceptionally broad and that have proven useful for structuring and stimulating research activity. Examples include basic terms such as attention, memory, or emotion (each of these notions can have hugely different connotations across subfields of psychology and neuroscience). Timbral taxonomies will need to be refined, depending on the subject matter. Importantly, researchers need to precisely specify which aspect or component of timbre they wish to address. The upshot of sharpened conceptual scalpels could be the development of refined experiments and more specific theories.

1.2 Milestones in Timbre Research

1.2.1 Fourier and Helmholtz

In 1863, Hermann von Helmholtz published the first edition of "On the Sensations of Tone as a Physiological Basis for the Theory of Music" (see Helmholtz 1877 for the English translation of the 4th German edition). The work was soon acknowledged as one of the most influential contributions to hearing science of the nineteenth century. Helmholtz's most important conceptual tool was Fourier's theorem. Providing a centerpiece of nineteenth century mathematics, Fourier conjectured that any periodic function can be represented as an infinite series of trigonometric functions. Ohm and Helmholtz applied the theorem to the description of sound and thereby demonstrated its usefulness for acoustic problems (Muzzulini 2006).

In practice, Fourier's theorem has led to the reduction of the infinite complexity of vibrational movements inherent in sounds to a finite number of parameters: the amplitudes and phases of a finite set of trigonometric functions, that is, a tone's partial components. This perspective also initiated the scientific study of timbre (for a comprehensive history of timbre research see Muzzulini 2006). Through experiments in sound synthesis and physiology, Helmholtz concluded that Fourier's theorem closely described physical and physiological reality. He used tuned resonators to filter out and amplify partial tones from a compound sound and concluded that the partial tones were physical entities that could be manipulated and experienced; they were not just mathematical fiction. With regard to physiology, he observed that "there must be different parts of the ear which are set in vibration by tones of different pitch [i.e., frequency] and which receive the sensation of these tones" (Helmholtz 1877, p. 143–144), thus providing the influential idea of the ear as a frequency analyzer (cf., Lyon 2017). Fourier analysis hence provided a common framework for the physics and physiology underlying auditory perception.

Regarding timbre, Helmholtz stated: "The quality of the musical portion of a compound tone depends solely on the number and relative strength of its partial simple tones, and in no respect on their difference of phase" (Helmholtz 1877, p. 126). This exclusively spectral perspective of timbre, locating the parameter in the relative amplitude of partial tones and nothing else, has dominated the field for a long time. But it is interesting to note how narrowly defined his object of study was, the "musical portion" of a tone: "… a musical tone strikes the ear as a perfectly undisturbed, uniform sound which remains unaltered as long as it exists, and it presents no alternation of various kinds of constituents” (Helmholtz 1877, p. 7–8). By assuming completely stationary sounds, his notion of tone color was indeed a strong simplification of what is understood as timbre today. Most obviously, attack and decay transients are not considered by this approach. Helmholtz was quite aware of this fact: “When we speak in what follows of a musical quality of tone, we shall disregard these peculiarities of beginning and ending, and confine our attention to the peculiarities of the musical tone which continues uniformly” (Helmholtz 1877, p. 67). This means that Helmholtz’s approach to timbre had its limitations (cf., Kursell 2013).

1.2.2 Timbre Spaces

Modern studies of timbre have started from direct dissimilarity ratings of pairs of sounds, a method that circumvents assumptions about acoustically important attributes and also does not rely on verbal descriptors. Multidimensional scaling (MDS) (Shepard 1962) has been a pivotal tool for this pursuit. Use of MDS generates a spatial configuration of points whose pairwise distances approximate the original perceptual dissimilarity data. In order to rule out potential confounds from other attributes, tones are usually equalized in pitch, loudness, and duration (and presented over headphones or a speaker, thereby removing any differences in spatial position) before entering a dissimilarity rating design. The central assumption of MDS studies is that shared psychophysical dimensions exist according to which the test sounds can be ordered. The goal of MDS studies is to reveal the dimensions that constitute the coordinate system of the timbre space.

The MDS approach has been an invaluable tool for modern timbre research. Although much of this work has traditionally revolved around musical-instrument sounds, MDS has also been applied in the scenarios of voice quality (Kreiman et al. 1992), industry product sounds and sound design (Susini et al. 2011), and timbre perception with cochlear implants (Marozeau and McKay 2016). The first application of MDS to timbre was provided by Plomp (1970) and Wessel (1973). In his dissertation, Grey (1975) used emulations of orchestral tones generated by means of additive synthesis with line-segment-approximated amplitude and frequency trajectories of partials extracted from analyses of musical-instrument tones. He observed a three-dimensional MDS solution. Its physical correlates were qualitatively interpreted in terms of the spectral energy distribution for the first dimension of the space. The second dimension was related to the attack synchronicity of partials, but sounds ordered along this dimension also had correspondingly different amounts of spectral fluctuation (variation over time). The third dimension was attributed to spectral balance during the attack of tones.

Using a set of sounds created by frequency-modulation synthesis, Krumhansl (1989) was the first to present a timbre space using EXSCAL (Winsberg and Carroll, 1989), an algorithm that includes so-called “specificities” that provide additional distance values to account for perceptual features that are unique to individual items. McAdams et al. (1995) synthesized many of the previously mentioned possibilities of MDS, including specificities with the addition of latent classes of subjects with different weights on the common dimensions and specificities using the CLASCAL algorithm (Winsberg and De Soete 1993) as well as rigorous quantification of physical correlates of the resulting MDS dimensions. Several audio descriptors were considered as candidates for a psychophysical interpretation of the MDS dimensions: *log rise time* (logarithm of the duration from the moment at which the start of the tone exceeds a certain threshold to the maximum amplitude), *spectral centroid* (amplitude-weighted mean frequency or center of mass of the spectrum), *spectral flux* (average of correlations between adjacent short-time amplitude spectra), and *spectral irregularity* (log of the standard deviation of component amplitudes of a tone's spectral envelope derived from a running average across the spectrum of the amplitudes of three adjacent harmonics).

Today, a number of MDS studies have confirmed that the spectral centroid and the attack time constitute major acoustic correlates of the MDS spaces from timbre dissimilarity ratings of orchestral musical-instrument sounds. The attack time appears to be particularly salient for stimulus sets that contain sustained and impulsively excited sounds, and additional dimensions appear to depend on the specific stimulus set. In this sense, these studies complemented the Helmholtzian approach by demonstrating that the temporal amplitude envelope is a salient timbral feature. At the same time, the low dimensionality of most of the obtained timbre spaces—usually studies observe around two to three dimensions—cast doubts with regards to their completeness. It is easy to imagine timbral variation that is not captured by these few dimensions, although these low-dimensional results may also reflect limitations in listeners' abilities to make ratings on more than a small number of perceptual factors simultaneously. The idea that musical-instrument timbre is indeed more complex is taken up by high-dimensional modulation representations (see Sect. 1.2.5).

1.2.3 Verbal Attributes

The plethora of words used to communicate timbral impressions of sounds further suggests a rich perceptual and conceptual dimensionality of timbre. Consider for example the following descriptions by Helmholtz:

> *Simple Tones* [single-frequency or flute-like sounds] … have a very soft, pleasant sound, free from all roughness, but wanting in power, and dull at low frequencies. … *Musical Tones* [piano- or vowel-like sounds] … are rich and splendid, while they are at the same time perfectly sweet and soft if the higher upper partials are absent. … If only the unevenly numbered partials are present, the quality of tone is hollow, and, when a large number of such upper partials are present, nasal. When the prime tone [fundamental] predominates, the quality of tone is rich; but when the prime tone is not sufficiently superior in strength to the upper

> partials, the quality of tone is poor. … When partial tones higher than the sixth or seventh are very distinct, the quality of tone is cutting and rough (Helmholtz 1877, pp. 118–119).

Soft, rough, wanting in power, dull, rich, sweet, hollow, nasal, poor, and cutting are just a few examples of the diverse and subtle lexicon of timbral attributes shared by instrumentalists, composers, sound engineers and designers, scientists, and other expert listeners, but also by naïve listeners who do not work with or study acoustics. These metaphorical descriptions are not crucial for *perceptualizing* timbre—one can compare, recognize, or memorize and imagine timbres without having to tag them verbally—but are central to *conceptualizing* timbre by allowing listeners to communicate subtle acoustic variations in terms of other, more commonly shared experiences, some of which are more sensory in nature, whereas others are more abstract and conceptual (Wallmark 2014). In other words, the way timbre is talked about can disclose significant information about the way it is perceived.

The advent of the semantic differential (SD) method (Osgood 1952) provided a powerful tool for empirical studies and models of the relation between the two. Semantic differentials are verbally anchored scales, typically constructed either by two opposing descriptive adjectives such as “bright-dull” or by an adjective and its negation as in “bright-not bright.” A set of sounds is judged against a relatively large number of such scales, which are then reduced to a small set of factors (dimensions explaining the most variance across all scales) and factor loadings (amount of variance in each scale explained by a factor). Similar to MDS studies, sound stimuli are usually equalized in pitch, loudness, and duration before entering a semantic rating design. Solomon (1958) first applied the SD approach to timbre, setting the stage for a rich tradition of research in timbre semantics from musical instruments to industrial product sounds (Carron et al. 2017).

Von Bismarck (1974) used synthetic spectra that mimicked vowels and instruments and empirically derived verbal scales (in German) suitable for describing such timbres (as opposed to a priori selection by the experimenter) and settled for a four-dimensional semantic space for timbre. The first dimension was defined by the differential scale *dull-sharp*, explained almost half of the total variance in the data, and correlated well with the spectral centroid. In an English experiment taking up some of Bismarck’s verbal scales but using dyads played from different wind instruments, Kendall and Carterette (1993) found that *dull-sharp* ratings were less stable, likely because sharp in English refers more often to pitch than to timbre. Convergent evidence from all subsequent studies in English (and in several other languages) corroborate the notion that a salient semantic dimension of timbre related to spectral energy distribution and concentration of energy in higher frequency bands is captured by the pair of polar adjectives *dull-bright*. Lichte (1941) had previously demonstrated empirically a correlation between dull-bright and the (constant) difference in amplitude between successive harmonic complexes (in principle this corresponds to a transposition of the spectral centroid).

The other dimensions found by von Bismarck were *compact-scattered*, *fullempty*, and *colorfulcolorless*, relating to notions of density, volume, and richness, respectively. Today most SD studies will yield a single dimension of fullness (or mass) that encompasses all such timbral impressions as well as a third common dimension of roughness (or texture) (Zacharakis et al. 2014). The three dimensions of brightness, roughness, and fullness correspond strongly, but not one-to-one, with three salient psychophysical dimensions along which listeners are known to perceive timbre similarity: duration of attack transient, midpoint of spectral energy distribution, and spectral variation or irregularity (Zacharakis et al. 2015). They also have been shown, in some cases, to be relatively stable across different languages (Zacharakis et al. 2014) and cultures (Alluri and Toiviainen 2012), although more systematic explorations would be necessary to establish a cross-cultural and language-invariant semantic framework for timbre.

1.2.4 Recognition of Sound Sources and Events

Although researchers have long been aware of timbre's role as a critical cue for sound recognition (McAdams 1993), the empirical exploration of this issue has really gained momentum only in the last 10 years. The importance of sound source categories and mechanics in the perception of musical-instrument timbre was first demonstrated by Giordano and McAdams (2010). In their meta-analysis of several timbre dissimilarity rating studies, same-family or same-excitation tones turned out to be rated similarly and tended to occupy similar regions of MDS spaces. These results indicated that significant associations between the perception of musical timbre and the mechanics of the sound source emerge even when not explicitly demanded by the task (also see Siedenburg et al. 2016b). Moreover, whereas working memory capacity for abstract and unfamiliar timbres is arguably rather low (Golubock and Janata 2013), general familiarity with timbres and the availability of corresponding sound source categories has been shown to improve timbre recognition from working memory (Siedenburg and McAdams 2017b).

An aspect that stands out across recognition studies is that the recognition of human voices is particularly fast and robust compared to other stimuli such as musical-instrument sounds. This may be intuitive from an evolutionary and ontogenetic point of view because the voice is a sound source with which all humans should be particularly familiar. Specifically, Agus et al. (2012) observed faster classifications of vocal sounds compared to sounds from percussion or string instruments. Suied et al. (2014) further observed that voices were more robustly recognized compared to other instrumental sounds even for very short snippets (below 10 ms duration). Extending this line of research toward recognition of musical melodies, Weiss and colleagues (see Weiss et al. 2017 and references therein) accumulated evidence for better recognition of vocal melodies compared to melodies played by nonvocal musical instruments.

How quickly can the sensory templates underlying sound-to-category mapping be acquired? Using fully abstract sounds, namely snippets of white noise, Agus et al. (2010) demonstrated that sensory representations are learned rapidly and are retained in fine-grained detail. Specifically, their experiment used short noise bursts, some of which re-occurred during the test unbeknownst to participants. Accuracy in the detection of repetitions embedded in noises itself increased rapidly for many of the repeated samples, and this type of implicit auditory learning turned out to be persistent over several weeks, which highlights the remarkable learning and recognition capabilities of the auditory system.

1.2.5 High-Dimensional Acoustic and Neuromimetic Representations

In speech processing and perception modeling, high-dimensional representations of audio signals have been common for some time (Dau et al. 1997; Chi et al. 1999). In this context, a debate revolves around the question of *how* "high-dimensional" the signal representations need to be in order to be able to parsimoniously account for the experimental data. The model developed by Dau et al. (1997) is based on a temporal-modulation filter bank but does not explicitly include information about spectral or spectrotemporal modulations. Directly inspired by physiological measurements of spectrotemporal receptive fields, Elhilali, Shamma, and colleagues (2003) have used a more complete set of spectrotemporal modulations in order to predict speech intelligibility. At the same time, for a task such as automatic speaker identification, it remains common practice to use fairly small sets of Mel-frequency cepstral coefficients (MFCC), which only represent spectral profile information of slices of the audio signal and hence no modulation information at all (Hansen and Hasan 2015).

In the field of music information retrieval, numerous studies have investigated robust timbre-related audio descriptors for tasks such as classification of orchestral instruments or music genres. In this context, researchers most often apply very large sets of hand-crafted audio descriptors (e.g., Siedenburg et al. 2016a). From a psychological viewpoint, this practice raises the question of the extent to which different acoustic descriptors are statistically independent of one another and whether they represent perceptually relevant information. Peeters et al. (2011) assessed the information redundancy across commonly used audio descriptors via correlational analysis followed by hierarchical clustering. This approach indicated ten classes of relatively independent acoustic descriptors. Applying receptive field models of auditory information processing to musical-instrument sounds, Patil et al. (2012) showed that robust, automatic instrument classification is possible on the basis of spectrotemporal modulation information, and Thoret et al. (2017) indicated that similar features are sufficient for characterizing the acoustic correlates of musical instrument identification.

A particularly useful trait of the representations used by Thoret and colleagues is that they are invertible, that is, they also can be used to generate sounds. This allows one to evaluate the importance of specific aspects of the underlying representations, which corresponds to the classic analysis-by-synthesis approach (Risset and Wessel 1999) (for applications to controlling the expressivity of musical performances, see Barthet et al. 2010). In the realm of sound texture perception, McDermott and Simoncelli (2011) presented an analysis-resynthesis scheme for texture exemplars such as rain, crashing waves, and wind, and had participants identify resynthesized signals. They found that by matching the statistics of individual frequency channels of the underlying auditory model, the approach failed to produce realistic resynthesized textures. By combining frequency channel statistics with correlations between channels, however, natural-sounding textures could be generated.

Audio-based models have thus started to become very useful tools to formulate hypotheses about the perceptual principles underlying timbre perception. The great diversity of approaches and representations across applications and signal classes that can be observed in the above examples may yet call for a revised understanding of the role of representations. Instead of seeking audio representations that act as repositories of everything that might be known about auditory information processing, audio-based models and representations can also be used pragmatically in order to support specific arguments about timbre perception (such as the importance of including cross-channel information). Useful insights are certain in the future from audio-based models and representations, which potentially may also be advanced by work with large-scale neural network models and analyses.

1.2.6 Neural Correlates of Timbre Processing

The emergence of functional magnetic resonance imaging (fMRI) has brought significant advances to the understanding of the physiological underpinnings of timbre perception by making it possible to nonintrusively measure correlates of brain activity in human listeners. Two general approaches to understanding the brain basis of timbre processing have been employed using different kinds of models. *Encoding models* are used to predict brain activity at the voxel level from stimulus properties. *Decoding models* attempt to predict stimulus properties from measurements of brain activity. Low-level representations of timbral properties examine the coding of spectral and temporal stimulus properties at different levels of auditory processing from the cochlea to auditory cortex and beyond. Spectral properties are represented by the distribution of activity across the tonotopic map at various levels (Town and Bizley 2013). Some temporal properties are presumably extracted by amplitude modulation filter banks, which are present as early as the inferior colliculus (Langner 2009), with evidence of a topography for rates of amplitude fluctuations in auditory cortex (Baumann et al. 2015).

Mid-level representations formed in secondary cortical areas capture descriptive summaries of sounds (such as roughness and brightness) that correspond to perceivable dimensions of timbre, and these properties then contribute to higher-level representations of sound sources. Early fMRI studies have demonstrated a distinct dorsal pathway for the processing of complex auditory patterns related to timbre

(Rauschecker 1998). That pathway provides information for the subsequent acquisition of knowledge of the environment and recognition of sound sources. Also using fMRI, Belin et al. (2000) found bilateral voice-selective areas in the superior temporal sulcus, part of the secondary auditory cortex. These areas, subsequently dubbed *temporal voice areas*, respond selectively to human vocal sounds but not to other sounds generated by humans or control sounds with matching amplitude envelopes.

Exploring facets of multimodal processing, von Kriegstein et al. (2005) reported robust interactions between auditory and visual areas during voice recognition. The authors found that brain regions involved in recognizing the voices of familiar speakers overlapped with the fusiform face area, a prominent face-sensitive region in the inferior temporal cortex. Several follow-up studies (see Mathias and von Kriegstein 2014) provided evidence of direct and early interactions between portions of the temporal voice areas and the fusiform face area, suggesting that these regions communicate with one another to resolve a speaker's identity.

Further evidence from fMRI studies suggests that processing related to the categorization of musical-instrument sounds, but not speech or animal vocalizations, occurs in the right superior temporal regions (Leaver and Rauschecker 2010). These authors also report other differences in localization of the processing of different classes of sounds: human speech and musical instruments versus animal vocalizations in anterior superior temporal cortex (STC) with preferential encoding of musical-instrument timbre in the right anterior superior temporal plane and selective processing of acoustic-phonetic content of speech in left STC.

Generally, this field is still very young. Methodological advances in the computational modeling of auditory perception (Kell et al. 2018) or the analysis of fMRI data (Diedrichsen and Kriegeskorte 2017) may well lead to a deeper understanding of the basis of timbre perception in the brain.

1.3 Structure and Content of Volume

1.3.1 Roadmap of Chapters

This volume is the first dedicated to a comprehensive and authoritative presentation of the state of the art in research on timbre. The first part addresses the principal processes underlying timbre perception and cognition and comprises five chapters. Chapter 2 by Stephen McAdams discusses dimensional models of timbre based on multidimensional scaling (MDS) of timbre dissimilarity ratings and psychophysical explanations in terms of acoustic correlates of perceptual dimensions. It covers research on the covariance of timbre, pitch, and loudness, and McAdams discusses the ways in which this covariance affects the recognition and identification of sound sources. Chapter 2 further discusses the utility of considering high-dimensional acoustic representations, such as modulation spectra, as an acoustic basis for timbre modeling.

Chapter 3 by Trevor Agus, Clara Suied, and Daniel Pressnitzer describes the many important and intriguing empirical findings on the categorization and recognition of sounds in the last 10 years or so. This chapter reviews these studies and specifically examines the minimal amount of acoustic and temporal information required to recognize sounds such as repeated noise bursts, isolated instrument sounds, or polyphonic musical textures. The chapter thus addresses the core question regarding the timbre cues utilized by humans for the recognition of various classes of sounds.

Chapter 4 by Kai Siedenburg and Daniel Müllensiefen discusses research on long- and short-term memory for timbre. A guiding question is whether timbre is stored independently from other mental tokens (e.g., pitch as in musical melodies or words as in verbal utterances) and whether it is governed by the same principles as those observed in these neighboring domains. Finding answers to these questions will involve decomposing memory for timbre into cognitive processes, such as perceptual similarity, chunking, and semantic encoding, as well as accounting for the factor of auditory expertise.

Chapter 5 by Charalampos Saitis and Stefan Weinzierl considers verbal descriptions of timbre and the rich semantic associations found in them. The authors look at how different communities of listeners verbally negotiate timbral qualities of sounds, the underlying conceptualizations of timbre, and the few salient semantic substrates. A critical question addressed is the relationship between the semantic and perceptual dimensions of timbre. To this end, acoustic correlates of verbal attributes and comparisons between semantic (language-based) and perceptual (dissimilarity-based) spaces of timbre are examined.

Chapter 6 by Vinoo Alluri and Sudarsana Reddy Kadiri reviews recent findings regarding the neural basis of timbre information processing from studies using both animal models and human brain imaging. This chapter addresses the specific neural correlates of spectral and temporal shape discrimination, findings regarding the cortical representation of spectrotemporal information, and more general models for the processing of sound source identity in cortex. Chapter 6 also examines the neural underpinnings of the perception of collections of timbres that characterize certain musical ensembles and composers.

The second part of this volume addresses specific scenarios of timbre perception. Chapter 7 by Samuel Mathias and Katharina von Kriegstein outlines important topics in voice processing and voice identification. Humans effortlessly extract a wealth of information from speech sounds, including semantic and emotional properties and details related to speaker identity. The chapter reviews the basic principles of human vocal production, behavioral studies on the processing and recognition of familiar and unfamiliar voices, as well as neural mechanisms and models of speaker recognition. The chapter further introduces phonagnosia, the deficit of not being able to recognize familiar people by their voices, and discusses its relation to autism spectrum disorder.

Chapter 8 by Stephen McAdams describes the various ways in which timbre shapes the perceptual experience of music. This chapter reviews the processes that may serve as the basis of this phenomenon with a particular focus on the principles of auditory scene analysis. Specific perceptual processes addressed include timbre's dependence on concurrent grouping (including timbral blend), the processing of sequential timbral relations, its role in sequential and segmental grouping, and the contribution of these grouping processes to musical structuring. The discussion draws from psychophysical studies and selected musical examples from the Western orchestral repertoire.

In Chap. 9, Jeremy Marozeau and Wiebke Lamping review timbre perception in patients with severe or profound hearing loss that have received a cochlear implant (CI). Although the perception of speech in quiet works relatively well for CI patients, music perception and voice identification still pose great problems. The chapter discusses CI research on timbre dissimilarity perception, musical instrument identification, and auditory stream segregation, issues in individual voice and gender recognition, and potential improvements for CI coding strategies.

Chapter 10 by Guillaume Lemaitre and Patrick Susini focuses on the role of timbre in the evaluation of product sounds, which is related to the question of how sounds contribute to the aesthetic, functional, and emotional aspects of a product. Research in this domain has utilized multidimensional scaling in conjunction with acoustic descriptor-based approaches and regression modeling in order to develop models of sound quality that can be applicable in sound design. Example cases of products are diverse: car horns, wind turbines, or consumer electronic devices such as printers. Implications for approaches to sonic interaction design are also discussed.

The third and final part of this volume is focused on the acoustic modeling of timbre. Chapter 11 by Marcelo Caetano, Charalampos Saitis, and Kai Siedenburg describes computational approaches to the acoustic description of sounds that have developed in the fields of psychoacoustics and music information retrieval to date. Having such tools at hand is essential for a better understanding of the psychological processes underlying the perception and cognition of timbre. Many scalar or time-varying descriptors are based on the Short-Time Fourier Transform from which summary measures are computed. Others are inspired by signal transformations that mimic physiological processes of audition.

Chapter 12 by Mounya Elhilali outlines recent advances in the study and application of spectrotemporal modulation representations in speech and music. This work has developed a neuro-computational framework based on spectrotemporal receptive fields recorded from neurons in the mammalian primary auditory cortex as well as from simulated cortical neurons. The chapter discusses the utility of applying this framework to the automatic classification of musical-instrument sounds and to robust detection of speech in noise.

Chapter 13 by Sølvi Ystad, Mitsuko Aramaki, and Richard Kronland-Martinet introduces an analysis-synthesis framework that derives intuitive control parameters of electronic sound synthesis directly from the statistics of input sounds. The framework is based on the distinction between action and object properties that are related to the mode of sound source excitation and resonance properties, respectively. The chapter reviews recent applications of this framework to the synthesis of impact sounds, textures, and musical-instrument sounds.

1.3.2 Future Perspectives

Although the thirteen chapters of this volume certainly lay out a wealth of information on timbre, research usually raises more questions than answers. In closing, a few words on promising directions for future work are in order. The following discussion is based on a query to the authors of this volume regarding the most important research topics of the next 10 years. The responses received have been condensed into roughly four main themes. Not surprisingly, these themes concern the foundations of timbre rather than some potential exotic extensions of the field:

(1) The chain of signal transformations from vibrations of physical bodies to brain signals is only poorly understood. Is sound source recognition based on the extraction (or pickup) of invariants (structural or transformational in Gibsonian terms) or on the learning of the covariation of various sensory properties (including those associated with timbre) across the many ways the object can be made to vibrate? More generally, how do the physics of the vocal tract or a musical instrument give rise to perceptually salient timbre features, how are these features processed in the brain, and how can knowledge about these principles lead to improved automatic sound source separation and recognition algorithms?

(2) Our understanding of timbre perception in everyday and musical contexts is still vague. Is it possible to establish a model of context-specific configurations of perceptual features that substantiates the current state of knowledge about timbre perception? Regarding the context of polyphonic music, is timbre a unitary percept or an emergent property of a multiplicity of percepts (drawing from pitch, the latter could be dubbed *Klangfarbenharmonie*)?

(3) How do the varieties of interindividual differences shape timbre perception? What may be good test batteries to compare the timbre perceptions of different individuals? The example of phonagnosia provides a fascinating window into this topic; however, even basic questions regarding differences between musicians and nonmusicians in basic timbre tasks have been explored only at a superficial level. Hearing impairment, our common fate, and its impact on timbre perception is yet another important interindividual factor that requires further exploration.

(4) Finally, what role does timbre, and particularly timbre-based expression, play in the communication of emotion and the evocation of emotion in the listener in speech and music? Closely related to this question is the need to specify the role of affective mediation in timbre semantics. Do verbal descriptions, such as bright versus dull, reflect perceptual or affective evaluation of sound qualities?

If the following chapters succeed in motivating future work on questions such as these, the goal of this volume would be fulfilled.

Acknowledgements KS was supported by a Carl von Ossietzky Young Researchers' Fellowship from the University of Oldenburg. CS wishes to thank the Alexander von Humboldt Foundation for support through a Humboldt Research Fellowship. SMc was supported by grants from the Canadian Natural Sciences and Engineering Research Council and Social Sciences and Humanities Research Council, the Fonds de recherche du Québec—Société et culture, the Canada Research Chairs Program, and the Killam Research Fellow Program of the Canada Council for the Arts.

Compliance with Ethics Requirements Kai Siedenburg declares that he has no conflict of interest.
Charalampos Saitis declares that he has no conflict of interest.
Stephen McAdams declares that he has no conflict of interest.

References

Agus TR, Thorpe SJ, Pressnitzer D (2010) Rapid formation of robust auditory memories: insights from noise. Neuron 66:610–618

Agus TR, Suied C, Thorpe SJ, Pressnitzer D (2012) Fast recognition of musical sounds based on timbre. J Acoust Soc Am 131(5):4124–4133

Alluri V, Toiviainen P (2010) Exploring perceptual and acoustical correlates of polyphonic timbre. Music Percept 27(3):223–241

Alluri V, Toiviainen P (2012) Effect of enculturation on the semantic and acoustic correlates of polyphonic timbre. Music Percept 29:297–310

ANSI (1960/1994) Psychoacoustic terminology: timbre, New York

Barthet M, Depalle P, Kronland-Martinet R, Ystad S (2010) Acoustical correlates of timbre and expressiveness in clarinet performance. Music Percept 28(2):135–153

Baumann S, Joly O, Rees A et al (2015) The topography of frequency and time representation in primate auditory cortices. eLife 4:03256. https://doi.org/10.7554/eLife.03256

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. Nature 403(6767):309–312

Bregman AS (1990) Auditory scene analysis: The perceptual organization of sound. The perceptual organization of sound. MIT Press, Cambridge

Carron M, Rotureau T, Dubois F et al (2017) Speaking about sounds: a tool for communication on sound features. J Design Res 15:85–109

Chi T, Gao Y, Guyton M et al (1999) Spectro-temporal modulation transfer functions and speech intelligibility. J Acoust Soc Am 106:2719–2732

Dau T, Kollmeier B, Kohlrausch A (1997) Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J Acoust Soc Am 102(5):2892–2905

Diedrichsen J, Kriegeskorte N (2017) Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Comp Bio 13(4):e1005508

Dolan EI (2013) The orchestral revolution: Haydn and the technologies of timbre. Cambridge University Press, Cambridge

Elhilali M, Chi T, Shamma SA (2003) A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Comm 41:331–348

Florentine M, Popper AN, Fay RR (eds) (2011) Loudness. Springer, New York

Giordano BL, McAdams S (2010) Sound source mechanics and musical timbre perception: evidence from previous studies. Music Percept 28(2):155–168

Golubock JL, Janata P (2013) Keeping timbre in mind: working memory for complex sounds that can't be verbalized. J Exp Psy: HPP 39(2):399–412

Grey JM (1975) An exploration of musical timbre. Dissertation, Stanford University

Hansen JH, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. IEEE Sig Proc Mag 32(6):74–99

Helmholtz H (1877) Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik, 4th edn. F. Vieweg und Sohn, Braunschweig. English edition: Helmholtz H (1954) On the sensations of tone as a physiological basis for the theory of music (trans: Ellis AJ), 2nd edn. Dover, New York

Kell AJE, Yamins DLK, Shook EN et al (2018) A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98:630–644. https://doi.org/10.1016/j.neuron.2018.03.044

Kendall RA, Carterette EC (1993) Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. Music Percept 10:445–468

Kreiman J, Gerratt BR, Precoda K, Berke GS (1992) Individual differences in voice quality perception. J Speech Lang Hear Res 35(3):512–520

Krumhansl CL (1989) Why is musical timbre so hard to understand? In: Nielzén S, Olsson O (eds) Structure and perception of electroacoustic sound and music. Excerpta Medica, Amsterdam, pp 43–53

Kursell J (2013) Experiments on tone color in music and acoustics: Helmholtz, Schoenberg, and Klangfarbenmelodie. Osiris 28:191–211

Langner G (2009) A map of periodicity orthogonal to frequency representation in the cat auditory cortex. Front Integr Neurosci 3:27. https://doi.org/10.3389/neuro.07.027.2009

Leaver AM, Rauschecker JP (2010) Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci 30:7604–7612. https://doi.org/10.1523/JNEUROSCI.0296-10.2010

Lichte WH (1941) Attributes of complex tones. J Exp Psychol 28:455–480

Lyon FL (2017) Human and machine hearing: extracting meaning from sound. Cambridge University Press, Cambridge

Marozeau J, McKay CM (2016) Perceptual spaces induced by Cochlear implant all-polar stimulation mode. Trends in Hearing 20. https://doi.org/10.1177/2331216516659251

Mathias SR, von Kriegstein K (2014) How do we recognise who is speaking? Front Biosci (Schol Ed) 6:92–109

McAdams S (1993) Recognition of sound sources and events. In: McAdams S, Bigand E (eds) Thinking in sound: the cognitive psychology of human audition. Oxford University Press, Oxford, pp 146–198

McAdams S, Winsberg S, Donnadieu S et al (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res 58(3):177–192

McDermott JH, Simoncelli EP (2011) Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. Neuron 71:926–940

Muzzulini D (2006) Genealogie der Klangfarbe (Geneology of timbre). Peter Lang, Bern

Osgood CE (1952) The nature and measurement of meaning. Psychol Bull 49:197–237

Patel AD (2008) Music, language, and the brain. Oxford University Press, Oxford

Patil K, Pressnitzer D, Shamma S, Elhilali M (2012) Music in our ears: the biological bases of musical timbre perception. PLoS Comp Biol 8(11):e1002759

Peeters G, Giordano BL, Susini P et al (2011) The timbre toolbox: audio descriptors of musical signals. J Acoust Soc Am 130:2902–2916. https://doi.org/10.1121/1.3642604

Plack CJ, Oxenham AJ, Fay RR, Popper AN (eds) (2005) Pitch. Springer, New York

Plomp R (1970) Timbre as a multidimensional attribute of complex tones. In: Plomp R, Smoorenburg GF (eds) Frequency analysis and periodicity detection in hearing. Suithoff, Leiden, pp 397–414

Rauschecker JP (1998) Cortical processing of complex sounds. Curr Opin Neurobiol 8(4):516–521. https://doi.org/10.1016/S0959-4388(98)80040-8

Risset J-C, Wessel DL (1999) Exploration of timbre by analysis and synthesis. In: Deutsch D (ed) The psychology of music, 2nd edn. Academic, San Diego, pp 113–169

Shepard R (1962) The analysis of proximities: multidimensional scaling with an unknown distance function. I. Psychometrika 27(2):125–140

Siedenburg K, McAdams S (2017a) Four distinctions for the auditory "wastebasket" of timbre. Front Psychol 8(1747)

Siedenburg K, McAdams S (2017b) The role of long-term familiarity and attentional maintenance in auditory short-term memory for timbre. Memory 25(4):550–564

Siedenburg K, Fujinaga I, McAdams S (2016a) A comparison of approaches to timbre descriptors in music information retrieval and music psychology. J New Music Res 45(1):27–41

Siedenburg K, Jones-Mollerup K, McAdams S (2016b) Acoustic and categorical dissimilarity of musical timbre: Evidence from asymmetries between acoustic and chimeric sounds. Front Psych 6(1977). https://doi.org/10.3389/fpsyg.2015.01977

Solomon LN (1958) Semantic approach to the perception of complex sounds. J Acoust Soc Am 30:421–425

Suied C, Agus TR, Thorpe SJ, Mesgarani N, Pressnitzer D (2014) Auditory gist: recognition of very short sounds from timbre cues. J Acoust Soc Am 135(3):1380–1391

Susini P, Lemaitre G, McAdams S (2011) Psychological measurement for sound description and evaluation. In: Berglund B, Rossi GB, Townsend JT, Pendrill LR (eds) Measurement with persons–theory, methods and implementation area. Psychology Press/Taylor and Francis, New York

Thoret E, Depalle P, McAdams S (2017) Perceptually salient regions of the modulation power spectrum for musical instrument identification. Front Psychol 8(587)

Town SM, Bizley JK (2013) Neural and behavioral investigations into timbre perception. Front Syst Neurosci 7:1–14. https://doi.org/10.3389/fnsys.2013.00088

von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud A-L (2005) Interaction of face and voice areas during speaker recognition. J Cogn Neurosci 17(3):367–376

von Bismarck G (1974) Timbre of steady tones: a factorial investigation of its verbal attributes. Acust 30:146–159

Wallmark Z (2014) Appraising timbre: embodiment and affect at the threshold of music and noise. Dissertation, University of California

Weiss MW, Schellenberg EG, Trehub SE (2017) Generality of the memory advantage for vocal melodies. Music Percept 34(3):313–318

Wessel DL (1973) Psychoacoustics and music: a report from Michigan State University. PACE: bulletin of the computer arts Society 30:1–2

Winsberg S, Carroll JD (1989) A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model. Psychometrika 54(2):217–229

Winsberg S, De Soete G (1993) A latent class approach to fitting the weighted Euclidean model, CLASCAL. Psychometrika 58(2):315–330

Zacharakis A, Pastiadis K, Reiss JD (2014) An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. Music Percept 31:339–358

Zacharakis A, Pastiadis K, Reiss JD (2015) An interlanguage unification of musical timbre: bridging semantic, perceptual, and acoustic dimensions. Music Percept 32:394–412

Part I
Principal Perceptual Processes

Chapter 2
The Perceptual Representation of Timbre

Stephen McAdams

Abstract Timbre is a complex auditory attribute that is extracted from a fused auditory event. Its perceptual representation has been explored as a multidimensional attribute whose different dimensions can be related to abstract spectral, temporal, and spectrotemporal properties of the audio signal, although previous knowledge of the sound source itself also plays a role. Perceptual dimensions can also be related to acoustic properties that directly carry information about the mechanical processes of a sound source, including its geometry (size, shape), its material composition, and the way it is set into vibration. Another conception of timbre is as a spectromorphology encompassing time-varying frequency and amplitude behaviors, as well as spectral and temporal modulations. In all musical sound sources, timbre covaries with fundamental frequency (pitch) and playing effort (loudness, dynamic level) and displays strong interactions with these parameters.

Keywords Acoustic damping · Acoustic scale · Audio descriptors · Auditory event · Multidimensional scaling · Musical dynamics · Musical instrument · Pitch · Playing effort · Psychomechanics · Sound source geometry · Sounding object

2.1 Introduction

Timbre may be considered as a complex auditory attribute, or as a set of attributes, of a perceptually fused sound event in addition to those of pitch, loudness, perceived duration, and spatial position. It can be derived from an event produced by a single sound source or from the perceptual blending of several sound sources. Timbre is a perceptual property, not a physical one. It depends very strongly on the acoustic properties of sound events, which in turn depend on the mechanical nature of vibrating objects and the transformation of the waves created as they propagate

S. McAdams (✉)
Schulich School of Music, McGill University, Montreal, QC, Canada
e-mail: stephen.mcadams@mcgill.ca

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_2

through reverberant spaces. The perceptual representation of timbre in the auditory system has been studied extensively. Such a representation is thought to underlie the recognition and identification of sound sources, such as human speech and musical instruments, or environmental events, such as rustling leaves, pounding surf, or a cooing dove.

Timbre encompasses a number of properties of sound events, such as auditory brightness (the mellowness of the horn versus the brightness of the muted trumpet), roughness (a growly jazz tenor saxophone), attack quality (sharp attack of a violin pizzicato versus the slow attack of a clarinet), hollowness (a clarinet sound), and inharmonicity (tubular bells). These properties also include traits that signal characteristics of the sounding body—its large or small size, geometry, and materials (wood versus metal)—and the way it was set into vibration (struck, blown, rubbed, rolled, and so on).

Essential questions that arise in studying timbre include the following:

- What perceptual representations of timbre are suggested by different behavioral and modeling approaches?
- To what extent are the modeled representations dependent on stimulus context?
- How does timbre interact or covary with pitch and loudness in acoustic sound sources?
- What differences are there between the role of timbre as a cue for the *identity* of a sounding object (including the action that sets it into vibration) and timbre's role as a *perceptual quality* that can be compared across separate events?

Certain aspects of timbre were studied as early as the late nineteenth century by Helmholtz (1885). He demonstrated that the "quality of sound," as Zahm (1892) (Caetano, Saitis, and Siedenburg, Chap. 11) refers to it, or *Klangfarbe* in the original German (literally "sound color"), is due to the number and relative intensity of the partials of a complex sound (i.e., its spectral envelope). For example, a voice singing a constant middle C while varying the vowel being sung can vary the shape of the sound spectrum independently of the perceived pitch and loudness. The seventeenth century concept of a sound being formed of different partial tones (Mersenne's law of the harmonics of a vibrating string) was instrumental in leading Helmholtz to this conception of timbre. Zahm (1892) claimed that Gaspard Monge (late eighteenth to early nineteenth century French mathematician) asserted that the quality of the sounds emitted by vibrating strings was due to the order and number of vibrations.

Exploration of the complex nature of timbre awaited the development of methodological tools, such as multidimensional scaling of dissimilarity ratings, developed in the 1950s and 1960s and first applied to timbre by Plomp (1970). However, real advances in the understanding of the perceptual representation of timbre required subsequent developments in musical sound analysis and synthesis. Wessel (1973) was probably one of the first to apply these developments to timbre and to demonstrate that the origins of timbre reside not only in spectral properties but in temporal properties as well. This approach led to the conception of timbre as a set of perceptual dimensions represented in a *timbre space*. However, some new concepts, partially derived from auditory neuroscience, are challenging this view by taking a more unitary approach in which timbre, rather than being a collection of individual properties, emerges from a complex higher-dimensional representation taken as a whole.

This chapter examines several aspects of the perceptual representation of timbre. Discrimination studies, multidimensional conceptions of timbre, the acoustic correlates of those dimensions, and complex spectromorphological conceptions of timbre are presented. The contributions of timbre to the perception of the geometry and materials of sound sources, and the actions that set them into vibration, are emphasized. The chapter also considers the interactions of timbre with other auditory attributes, such as pitch and loudness, and playing effort of a musical instrument.

2.2 Timbre Discrimination

Discrimination performance is measured for sounds that have been modified in some way to determine which modifications create significant perceptual effects. There are few studies of the discrimination of specific timbre-related acoustic parameters. A study of the discrimination of linear rise and decay times in 1 kHz sine tones and noise bursts found that the just noticeable difference was about 25% of the duration of the rise or decay time, but discrimination was a bit better at times above 80 ms and much worse at times below 20 ms (van Heuven and van den Broecke 1979). Discrimination of decay times in noise bursts was best at moderate values, whereas rise times of sine tones were best discriminated at very short times when energy splatter probably provided a cue.

Experiments on discrimination of musical-instrument tones have often progressively simplified the sounds. One kind of simplification involves performing a fine-grained acoustic analysis of instrument tones and then resynthesizing them with modifications. Grey and Moorer (1977) presented listeners with different versions of string, woodwind, and brass tones: the original recorded tones and resynthesized versions of each one with various kinds of modifications (Fig. 2.1). These experiments showed that simplifying the pattern of variation of the amplitudes and frequencies of individual components in a complex sound affected discrimination for some instruments but not for others. When the attack transients (low-level noisy components at the very onset of the signal; see Fig. 2.1b) were removed, the tones were easily discriminated from the originals. Applying the same amplitude variation to all of the components (thus replacing the individual variations normally present) grossly distorted the time-varying spectral envelope of the tone and was easily discriminated. Complete removal of frequency change during the tone was also easily discriminated, although applying a common frequency variation to all components had only a weak effect on discriminability. These findings demonstrate a fine-grained perceptual sensitivity to the spectrotemporal microstructure of sound events.

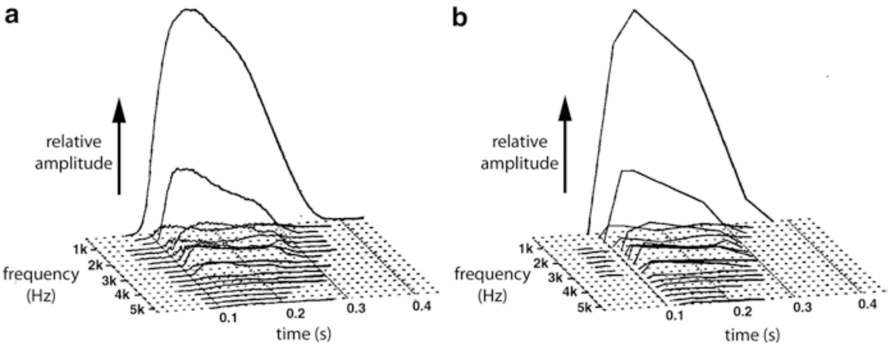

Fig. 2.1 Analysis of the time-varying amplitudes and frequencies of the partials of a bass clarinet tone (**a**) and their simplification by line segment functions (**b**). In this three-dimensional representation, *time* goes from left to right, *relative amplitude* from bottom to top, and *frequency* from back to front. Each curve shows the frequency and amplitude trajectory of a partial in the tone. Note the low-level inharmonic partials at the beginning of the sound, which are called *attack transients*. Attack transients are present in many sustained sounds and indicate the chaotic behavior of the sound coming from the instrument before it settles into a periodic vibration (Reproduced from figures 2 and 3 in Grey and Moorer 1977; used with permission of The Acoustical Society of America)

Similar results were obtained with more fine-grained modifications by McAdams et al. (1999). Spectral analyses of sounds from several instruments were used to produce time-varying harmonic amplitude and frequency representations that were then simplified in several ways and resynthesized. Listeners had to discriminate a reference sound resynthesized with the full data from a sound transformed with from one to four simplifications, which affected the amplitude and frequency behavior of the harmonics and the overall spectral envelope (the general shape of the amplitudes of the partials over frequency). Discrimination between the original reference sound and the various simplified sounds was very good when the spectral envelope was smoothed out and when the component amplitudes were made to vary together rather than independently. However, discrimination was moderate to poor when the frequency behavior of the partials was modified or the amplitude envelopes of the individual partials were smoothed. Discrimination of combinations of simplifications was equivalent to that of the most discriminable simplification. Analysis of the spectral data for changes in harmonic amplitude, changes in harmonic frequency, and changes in the "center of gravity" of the frequency spectrum (the amplitude-weighted mean frequency, more simply referred to as the *spectral centroid*) resulting from the simplifications revealed that these measures correlated well with discrimination results, indicating yet again that listeners have access to a relatively fine-grained sensory representation of musical-instrument sounds.

One difficulty in generalizing these results to everyday situations is that perception of isolated tones may differ from that of tones in musical sequences. To test the effect of sequences on timbre discrimination, Grey (1978) used the same kind of simplified tones from Grey and Moorer (1977) for three instruments (bassoon, trumpet, and clarinet). He created notes at other pitches by transposing the instrument spectrum to higher or lower frequencies. Listeners were asked to discriminate between the original stimulus and the simplifications of a given instrument for either isolated tones or for the same tones placed in musical patterns that differed in rhythmic variety, temporal density, and number of simultaneous melodic lines. An increasingly complex musical context (isolated tones versus sequences) did not affect discrimination between original and modified versions of the bassoon but hindered such discrimination for the clarinet and trumpet. Small spectral differences were slightly enhanced in single-voice contexts compared with isolated tones and multi-voiced contexts, although discrimination remained high. Articulation differences, on the other hand, were increasingly disregarded as the complexity and density of the context increased. These results suggest that in cases where demands on perceptual organization and the storing and processing of sequential patterns are increased, fine-grained temporal differences are not preserved as well as spectral differences.

One possible confounding factor in Grey's (1978) study is that the different pitches were created by transposing a single tone's spectrum and then concatenating and superimposing these tones to create the musical patterns. This removes any normal variation of spectral envelope with pitch as well as any articulation features that would be involved with passing from one note to another in a melody. Kendall (1986) controlled for these problems in an instrument recognition experiment in which the recorded melodic sequences were modified by cutting parts of the tones and splicing them together. Listeners had to decide which of the instruments (clarinet, trumpet, or violin) playing an unedited melody matched the one playing the melody composed of modified sounds. Modifications of the normal tones included cutting attacks and decays (thereby leaving only the sustain portion) and presenting transients only (with either a silent gap in the sustain portion or an artificially stabilized sustain portion). The results suggest that transients in isolated notes provide information for instrument recognition when alone or coupled with a natural sustain portion but are of little value when coupled with a static sustain part. They are also of less value in continuous musical phrases in which the information present in the sustain portion (most probably related to the spectral envelope) is more important.

From these studies on the effects of musical context on discrimination, it can be concluded that the primacy of attack and legato transients found in all of the studies on isolated tones is greatly reduced in whole phrases (particularly slurred ones). The spectral envelope information present in the longer segments of the sustain portion of musical sounds is thus of greater importance in contexts where temporal demands on processing are increased.

2.3 Multidimensional Conceptions of Timbre

Dissimilarity ratings can be used to discover the salient dimensions that underlie the perception of a set of sounds. All possible pairs from the set are presented to a listener who rates how dissimilar they are on a given scale (say 1–9, where 1 means identical or very similar and 9 means very dissimilar on a continuous scale). In multidimensional scaling (MDS), the ratings are treated as psychological proximities between the judged items, and a computer program maps the dissimilarity ratings onto a spatial configuration in a given number of dimensions. The resulting geometrical structure is interpreted as reflecting the perceptual qualities listeners used to compare the sounds. In order to give a psychoacoustic meaning to the spatial representation, the dimensions of the space are correlated with acoustic properties of the tones. It is presumed that the dimensions on which listeners do focus are determined *firstly* by the set of sounds used in the experiment, that is, their representations may be coded with respect to the stimulus context provided within an experimental session, and *secondly* by knowledge or previous experience that listeners have with the classes of sounds used. In sum, this approach aims to give us an idea of the auditory representations that listeners use in comparing sounds.

One methodological advantage of the MDS approach is that listeners don't have to focus on a specific property to be rated, which has to be communicated to them with words that in turn are often ambiguous with respect to their meaning (but see Saitis and Weinzierl, Chap. 5). They simply rate how dissimilar all pairs of a set of sounds are (for reviews see Hajda et al. 1997; McAdams 2013).

2.3.1 Multidimensional Scaling Models

MDS routines compute a model of the dissimilarities in terms of Euclidean distances among all pairs of sounds in a stimulus set. The result is a space with a small number of shared perceptual dimensions. Various techniques are used to decide on the dimensionality of the model, some more qualitative, like stress values, and some more statistically based, like the Bayesian Information Criterion and Monte Carlo testing (for more detail see McAdams et al. 1995).

The basic MDS algorithm originally developed by Kruskal (1964) is expressed in terms of continuous dimensions that are shared among stimuli. The underlying assumption is that all listeners use the same perceptual dimensions to compare them. The model distances are fit to the empirically derived proximity data (usually dissimilarity ratings or confusion ratings among sounds). More complex algorithms like EXSCAL also include specificities (properties that are unique to a sound and increase its distance from all the other sounds beyond the shared dimensions), whereas others include different perceptual weights accorded to the dimensions and specificities by individual listeners (INDSCAL) or by latent classes of listeners (CLASCAL). The equation defining distance in the more general CLASCAL model (McAdams et al. 1995) is:

$$\partial_{ijc} = \sqrt{\sum_{D}^{d=1} w_{cd}\left(x_{id} - x_{jd}\right)^{2} + v_{c}\left(s_{i} + s_{j}\right)} \tag{2.1}$$

where ∂_{ijc} is the distance between sounds i and j for latent class c; x_{id} is the coordinate of sound i on dimension d; D is the total number of dimensions; w_{cd} is the weight on dimension d for class c; s_i is the specificity on sound i; and v_c is the weight on the whole set of specificities for class c. The basic MDS algorithm doesn't model weights or specificities and only has one class of listeners. EXSCAL has specificities, but no weights. INDSCAL has no specificities but has weights on each dimension for each listener.

One of the difficulties of the paired-comparison approach is that the number of dissimilarity ratings that each listener has to make increases quadratically with the number of sounds to be compared. To get around this limitation, Elliott et al. (2013) used the SMACOF algorithm to perform multiway constrained MDS in which multiple similarity ratings from different listeners are used for each pair of stimuli. In this paradigm a given listener only has to rate a subset of a large set of stimulus pairs.

2.3.2 Timbre Spaces

The result of an analysis applied to dissimilarity ratings of musical sounds of similar pitch, duration, and loudness is a timbre space, which characterizes the perceptual dimensions shared by a set of sounds. One underlying assumption is that the perceptual dimensions are orthogonal and should be characterizable by independent physical properties.

The most cited timbre space is from the seminal study by Grey (1977), using sustained musical-instrument sounds (blown and bowed) that had been analyzed and then resynthesized in simplified form (as in Fig. 2.1b). Using INDSCAL, he found a space with three dimensions (Fig. 2.2a). The first dimension corresponded qualitatively with the *spectral energy distribution*: brighter or more nasal sounds were at one extreme and mellower sounds were at the other. The second dimension was related to the degree of spectral fluctuation during the sound and the onset synchrony of harmonics (what has subsequently come to be called *spectral flux* or *spectral variation*). The position of sounds along the third dimension seemed to depend on the strength of attack transients, which characterizes the attack quality. Grey and Gordon (1978) validated the interpretation of the spectral dimension by exchanging the spectral envelopes of four pairs of sounds among the sixteen original ones that differed primarily in terms of this dimension (sounds connected by lines in Fig. 2.2). For example, the spectral envelope of the trumpet sound was applied to the muted trombone and vice versa. When they ran the study on this modified set, the pairs with switched spectral envelopes also switched positions along this dimension, confirming the interpretation (Fig. 2.2b).

It is important to note that although some features related to spectral distribution and temporal envelope seem ubiquitous (at least in musical sounds), the actual dimensions found depend on the type of acoustic variation that is present in the set

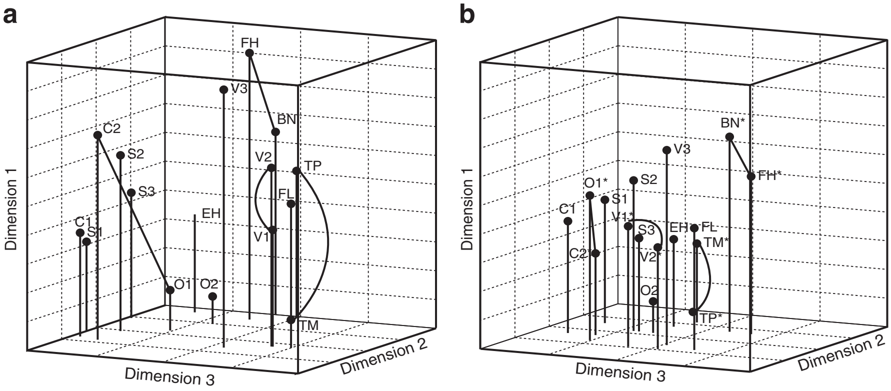

Fig. 2.2 Timbre spaces: (**a**) from Grey (1977) and (**b**) from Grey and Gordon (1978). *Dimension 1* is related to the spectral envelope distribution. *Dimension 2* corresponds to the amount of fluctuation over time in the spectral envelope and the synchrony of onset of the harmonics. *Dimension 3* captures the strength of attack transients. *BN*, bassoon; *C1*, Eb clarinet; *C2*, bass clarinet; *EH*, English horn; *FH*, French horn; *FL*, flute; *O1*, oboe 1; *O2*, oboe 2; *S1*, alto saxophone playing *piano*; *S2*, alto saxophone playing *mezzoforte*; *S3*, soprano saxophone; *TM*, trombone with mute; *TP*, trumpet; *V1*, violoncello playing *normale; V2*, violoncello playing *sul tasto* with mute; *V3*, violoncello playing *sul ponticello* (Modified from figures 2 and 3 in Grey and Gordon 1978; used with permission of The Acoustical Society of America)

of sounds being tested. The first timbre dissimilarity study to include percussion sounds was conducted by Lakatos (2000). He presented different sound sets to listeners: one with harmonic wind and string sounds (sustained and impulsive), one with percussion sounds (some pitched, like vibraphone or temple block, and some unpitched, like snare drum and cymbal), and a third one with ten sounds from each of those sets. A reanalysis of these data by McAdams (2015) found two dimensions for the wind/string set that qualitatively included spectral envelope and temporal envelope; those for the percussion set included temporal envelope and either spectral density or pitch clarity/noisiness of the sound. The combined set had all three: spectral distribution, temporal envelope, and spectral density.

One might wonder how much the relations among sounds, as determined by the dissimilarity ratings, depend on the global stimulus context. For example, if one were to change some of the sounds in a stimulus set or add new sounds that are quite different, would the relations among the original sounds be distorted, perhaps due to making the listener focus on different sound properties? In the reanalysis of Lakatos' (2000) dissimilarity data, McAdams (2015) compared the perceptual structure of the ten sounds from the wind/string and percussion sets that were included in the combined space with their structure in the original sets. With the exception of one percussion instrument, the relations among the ten sounds of each set maintained their dissimilarity relations in the presence of the very different new sounds from the other set. This result is important in demonstrating a relative robustness of timbre relations across different orchestration contexts. How would this apply in a musical setting? If, for instance, part of a piece uses the differences between string and woodwind instruments, listeners will tune in to the resulting timbral relations. If the composer then adds brass and percussion at a different point, these perceptual relations among string and woodwind sounds won't necessarily be perturbed by the new orchestral context.

The apparent assumption that extremely complex sounds like musical-instrument tones differ in terms of only a few common perceptual dimensions is questioned by many musicians. Each instrument may also produce unique characteristics that are not easily coded along a continuous dimension, such as the sudden pinched offset of a harpsichord, the odd-harmonic structure of the clarinet spectrum, or the amplitude modulation of a flutter-tongued flute or trumpet. Krumhansl (1989) used a set of sounds created by digital sound synthesis that imitated some musical instruments or that were conceived as hybrids of instruments, so the *guitarnet* was a chimera with the "head" of a guitar and the "tail" of a clarinet. An MDS analysis with EXSCAL produced a three-dimensional space with specificities. The analysis of specificities showed that a significant amount of variability in the similarity judgements, which could not be attributed to the common dimensions, could be accounted for by postulating unique features for some of the instruments, such as the simulated harp, harpsichord, clarinet, and vibraphone. This technique seems promising for identifying sounds that have special perceptual features, but it remains tricky to tie them to specific acoustic properties given that they are unique for each instrument.

Algorithms such as INDSCAL and CLASCAL allow for differences among individual listeners or latent classes of listeners, respectively. These differences are modeled as weighting factors on the different dimensions for both algorithms and on the set of specificities for CLASCAL. Latent classes are formed of listeners having a similar weight structure in their data. For example, one group of listeners might pay more attention to spectral properties than to temporal aspects, whereas another group might have the inverse pattern. McAdams et al. (1995) found five classes in a set of 84 listeners. Most of the listeners were in two classes that had fairly equal weights across dimensions and specificities. They merely differed in that one class used more of the rating scale than the other. For the other three classes, some dimensions were prominent (high weights) and others were perceptually attenuated (low weights). However, an attempt to link the classes to biographical data, including the amount of musical experience or training, was not conclusive. McAdams et al. (1995) found that similar proportions of nonmusicians, music students, and professional musicians fell into the different latent classes. One explanation may be that because timbre perception is so closely allied with the ability to recognize sound sources in everyday life, everybody is an expert to some degree, although different people are sensitive to different features.

Along the same lines of thought, the previously mentioned robustness of timbre spaces to changes in stimulus context may be due to the fact that timbre perception is strongly related to the recognition and categorization of sound sources (also see Agus, Suied, and Pressnitzer, Chap. 3). To test this idea, Giordano and McAdams (2010) conducted a meta-analysis of previously published data concerning identification rates and dissimilarity ratings of musical-instrument tones. The aim was to ascertain the extent to which large differences in the mechanisms for sound production (different instrument families, for example) were recovered in the perceptual data. In the identification studies, listeners frequently confused tones generated by musical instruments with a similar physical structure (e.g., clarinets and saxophones are often confused, both being single-reed instruments), but they seldom confused tones generated by very different physical systems (e.g., one rarely mistakes a trumpet, a lip-valve instrument, for a bassoon, a double-reed instrument, and never for a vibraphone, a struck metal bar). Consistent with this hypothesis, the vast majority of previously published timbre spaces revealed that tones generated with similar resonating structures (e.g., string instruments versus wind instruments) or with similar excitation mechanisms (e.g., impulsive excitation as in violin pizzicati versus sustained excitation as in a flute tone) occupied the same region in the space. To push this idea even farther, Siedenburg et al. (2016) presented recorded musical-instrument sounds previously determined to be highly familiar to listeners and digitally transformed versions of these sounds rated as highly unfamiliar. The dissimilarity ratings demonstrated that similarity between the source/cause mechanisms can affect perceived similarity, thereby confirming the meta-analysis results of Giordano and McAdams (2010).

As mentioned in Sect. 2.1, timbre emerges from the perceptual fusion of acoustic components into a single auditory event. This includes the perceptual fusion of sounds produced by separate instruments into a single blended event, a technique often used by instrumental composers to create new timbres (see McAdams, Chap. 8 for more on timbral blend). One question that arises concerns the extent to which the timbral properties of a blended event can be determined by the constituent events. Kendall and Carterette (1991) recorded dyads of different wind instruments that performed together. The dyads were presented to listeners who rated the dissimilarities between them. They found that the relations among dyads could be modeled as a quasi-linear combination of the positions of the individual instruments in timbre space. That is, if one determines the vector between two instruments (e.g., flute and saxophone) in a timbre space, the position of the flute/saxophone dyad would be at the point of bisection of that vector. This result suggests that in the case of dyads, there may not be much partial masking of the sound of one instrument by that of the other. However, one might imagine that this would begin to break down for blends of three or more instruments as the combined frequency spectrum densifies and auditory masking increases.

One last issue with the notion of timbre space is the degree to which the dimensions, which are modeled as orthogonal, are actually perceptually independent. Caclin et al. (2007) created synthesized harmonic sounds that varied independently in spectral centroid (see Sect. 2.4.1), attack time, and the ratio of the amplitudes of even and odd harmonics (related to the hollow quality of the clarinet). To explore the interaction of these dimensions, they employed a task in which dimensions were paired, and two values along each dimension were chosen so that the relative change along the two dimensions is equivalent, for instance, slow and fast attack versus bright and dull spectral envelope. Listeners were asked to focus on changes along only one of the dimensions and to ignore changes along the other. They had to categorize the sounds as quickly as possible along the criterial dimension. In one test, there was no change on the irrelevant dimension (called the baseline), and in others the sounds varied randomly, congruently (sharper attack and brighter timbre), or incongruently (sharper attack and mellower timbre) along the dimension to be categorized. If there is a cost in terms of speed and accuracy of categorization (i.e., it slows the listener down to have to ignore a change in attack when judging brightness and they make more errors), then the dimensions are considered to interact. This was the case for all three pairs of dimensions. So although these same three dimensions have fairly separate neural representations in auditory sensory memory (Caclin et al. 2006), the perceptual interaction supports a model with separate processing channels for those dimensions but with crosstalk between the channels.

2.3.3 Acoustic Correlates of Timbre Space Dimensions

Once a timbre space is obtained, the next stage in the psychophysical analysis is to determine the physical properties that determine the nature of the different dimensions. The primary approach is to define parameters derived from the audio signal that are strongly correlated with the position along a given perceptual dimension for a specific sound set. Grey and Gordon (1978) proposed the spectral centroid as a scalar correlate of the position of sounds along their spectral-envelope-related dimension. McAdams et al. (1995) were perhaps the first to try computing acoustic descriptors correlated with each perceptual dimension in a timbre space. For their three-dimensional space representing 18 synthetic sounds created with frequency-modulation synthesis, they found strong correlations between the position along the first dimension and attack time (Fig. 2.3) and between the position along the second dimension and the spectral centroid (Fig. 2.4). There was a weaker correlation between the position along the third dimension and the degree of variation of the spectral envelope over the duration of the tones (Fig. 2.5).

Subsequently, two major toolboxes with a plethora of quantitative descriptors were developed: the MIR Toolbox of Lartillot and Toiviainen (2007) and the Timbre Toolbox of Peeters et al. (2011) (although some of the timbre-related descriptors in both toolboxes have been criticized by Kazazis et al. 2017 and Nymoen et al. 2017). Some of the descriptors are derived from spectral properties, such as the first four moments of the frequency spectrum (centroid, spread, skew, kurtosis), measures of spectral slope, or the jaggedness of the spectral envelope. Other descriptors are derived from the temporal envelope, such as attack time and decay time. Still others capture time-varying spectral properties, such as spectral flux, a scalar value that represents the variability of the spectrum over time. Chapter 11 (Caetano, Saitis, and Siedenburg) provides more details on audio descriptors for timbre.

In many attempts to model timbre, authors have often chosen descriptors that seem most relevant to them, such as the spectral centroid (related to timbral brightness or nasality), attack time of the energy envelope, spectral variation or flux, and spectral deviation (jaggedness of the spectral fine structure). These vary from study to study making it difficult to compare results across them. Furthermore, many

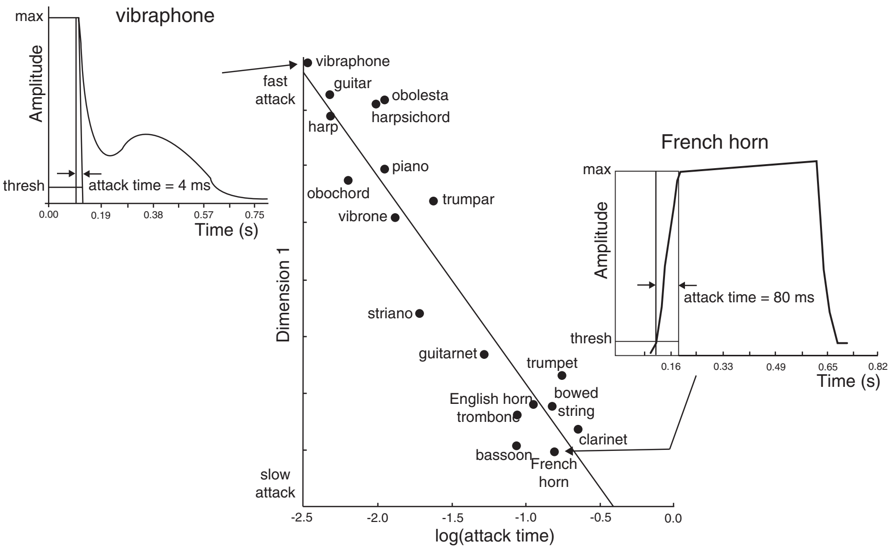

Fig. 2.3 Relationship of log (attack time) to position along Dimension 1. The diagrams for vibraphone and French horn show the *global amplitude envelope* (amplitude variation over time). The *attack time* was measured as the time from a threshold value to the maximum in the amplitude envelope. The attack time is much quicker for an impacted metal bar (*vibraphone*) than for a sustained wind instrument (*French horn*). The sounds were imitations of musical instruments or hybrids of instruments produced with frequency-modulation synthesis. The hybrids were the *guitarnet* (guitar/clarinet), *obochord* (oboe/harpsichord), *obolesta* (oboe/celesta), *striano* (bowed string/piano), *trumpar* (trumpet/guitar), and *vibrone* (vibraphone/trombone). *Sampled pian*o imitates an electronically sampled piano (Modified from figure 5 in McAdams 2013; used with permission from Elsevier)

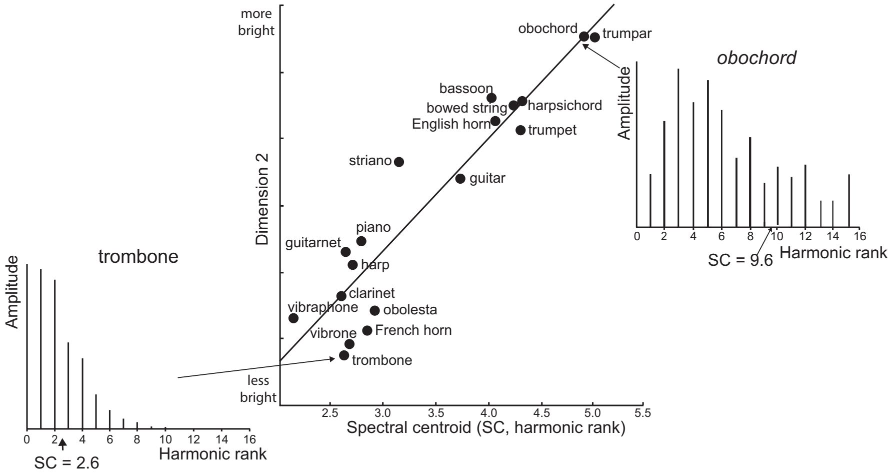

Fig. 2.4 Relationship of the spectral centroid to position along Dimension 2. The diagrams for *trombone* and *obochord* show the frequency spectra. The balance point in the energy spectrum (*spectral centroid, SC*) for each is shown. The SC is lower for trombone, which has much less energy in the higher harmonics, than for *obochord*, which has a rich spectrum with prominent higher harmonics (Modified from figure 4 in McAdams 2013; used with permission from Elsevier)

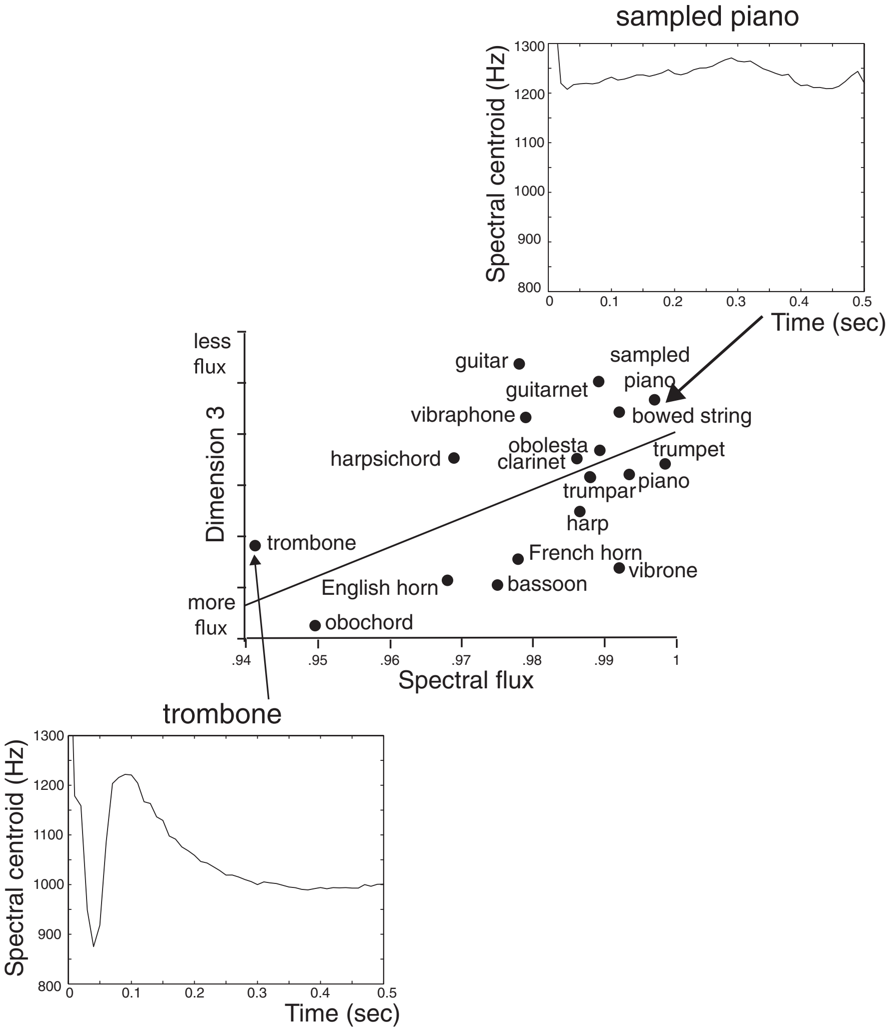

Fig. 2.5 Relationship of spectral flux to position along Dimension 3. The measure of *spectral flux* is the average correlation of the spectra between adjacent time frames. *Less flux* gives higher correlation values and *more flux* gives lower values. The diagrams for trombone and sampled piano show the variation of the spectral centroid (*SC*) over time. It is clear that there is more variation for trombone than for sampled piano (Reproduced from figure 6 in McAdams 2013; used with permission from Elsevier)

groups of descriptors capture similar spectral, temporal, or spectrotemporal properties and may not be independent of one another. To address this issue, Peeters et al. (2011) computed several measures on a set of over 6000 musical-instrument sounds with different pitches, dynamic markings (*pp* is very soft, *ff* is very loud), and playing techniques. These measures included the central tendency (median) and variability over time (interquartile range) of the time-varying acoustic descriptors in the Timbre Toolbox, as well as global scalar descriptors derived from the temporal

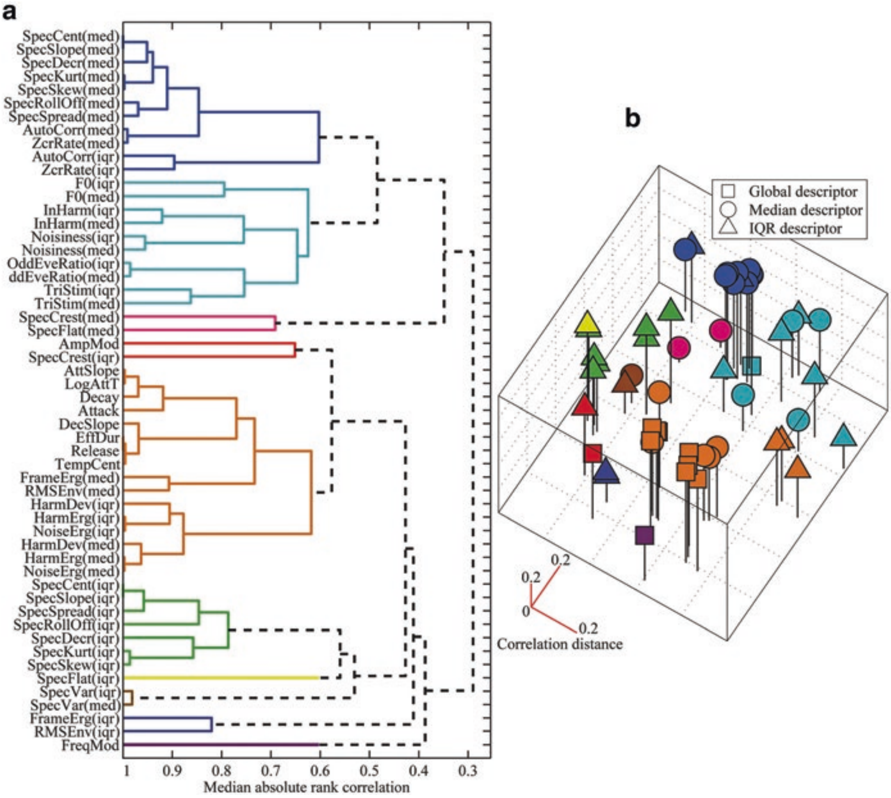

Fig. 2.6 Structure of similarities among audio descriptors. (**a**) The results of a hierarchical cluster analysis of the correlations among the audio descriptors listed along the y axis. Scalar values derived from the temporal energy envelope cluster in the middle. Statistical measures of time-varying descriptors include the median (*med*) as a measure of central tendency and the interquartile range (*iqr*) as a measure of variability. Different *colors* are used to highlight different clusters of descriptors. (**b**) A three-dimensional MDS (multidimensional scaling) of the between-descriptor correlations. Descriptors that are similar will be close in the space. The same *color scheme* is used in both panels to demonstrate the similarity of groups of descriptors. (Reproduced from figure 4 in Peeters et al. 2011, refer to that paper for more detail on the audio descriptors; used with permission of The Acoustical Society of America)

energy envelope. They found that many of the descriptors covaried quite strongly within even such a varied set of sounds. Using a hierarchical cluster analysis of correlations between descriptors over the whole sound set, they concluded that there were only about ten classes of independent descriptors (Fig. 2.6). This can make the choice among similar descriptors seem rather arbitrary in some cases, and just putting all available descriptors into a regression or other kind of model may seriously overfit the data.

No studies of timbre similarity have employed an approach in which the time-varying spectral properties are used as a time series, which may be intimately tied to both the mechanical nature of the sounding object and the way it is set into vibration. The domain of multi-objective time-series matching in which several time-varying properties are used collectively to measure similarity among sounds or for audio classification may show a way forward (Esling and Agon 2013).

The chaotic proliferation of audio descriptors in timbre research and in music information retrieval has seldom asked the question of whether these descriptors (or combinations of them) actually correspond to perceptual dimensions. Are they ordered on ordinal, interval, or ratio scales? To what extent are they perceptually independent? One confirmatory MDS study makes a small step in this direction. Caclin et al. (2005) analyzed dissimilarity ratings on purely synthetic sounds in which the exact nature of the stimulus dimensions could be controlled. These authors confirmed that perceptual dimensions related to the spectral centroid, log attack time, and spectral deviation (jaggedness of the spectral envelope) are orthogonal and demonstrated that they can at least be considered as interval scales. However, they did not confirm spectral flux, which seems to collapse in the presence of an equivalent perceptual variation in the spectral centroid and attack time. Another question concerns whether perceptual dimensions might actually arise from linear or nonlinear combinations of descriptors that are learned implicitly from long-term experience of their covariation in environmental, musical, and speech sounds. Stilp et al. (2010) demonstrated that a passive exposure to highly correlated acoustic properties leads to implicit learning of the correlation and results in a collapse of the two unitary dimensions (temporal envelope and spectral shape in their case) into a single perceptual dimension.

A number of studies have focused on the perceptual dimension correlated with the spectral centroid (often referred to as timbral brightness; see Saitis and Weinzierl, Chap. 5). Schubert and Wolfe (2006) compared two models of brightness: the spectral centroid (in units of Hz) and the centroid divided by the fundamental frequency (in units of harmonic rank). Listeners compared digital samples of two instruments (less bright piccolo, brighter trumpet) played at different pitches (E2, E4, A#4, E5; where C4 is middle C with a fundamental frequency of 261.6 Hz.) and dynamics (forte, piano). They were asked to rate the brightness, pitch, and loudness differences. Brightness ratings scaled better with the raw spectral centroid than with the fundamental-adjusted (and pitch-independent) centroid. It should be noted that timbre covaries strongly with both fundamental frequency and playing effort in acoustical instruments (see Sect. 2.6). Furthermore, a brightness model scaled for fundamental frequency would only be applicable to harmonic sounds.

From the same research group, another study examined ratio scaling of timbral brightness by adjusting the spectral slope of a synthesized sound to make it twice as bright as a reference sound (Almeida et al. 2017). They found that the ratio of spectral centroids to double the brightness was about 2.0 on average for a reference centroid of 500 Hz and decreased to about 1.5 for a reference centroid of 1380 Hz. This result suggests that timbral brightness is indeed a perceptual dimension that forms a ratio scale.

Finally, Siedenburg (2018) confirmed that shifts in spectral maxima are perceived as changes in brightness. His study also presents a timbral analogy to Shepard's (1964) pitch-circularity illusion in which the heights of local spectral peaks conform to a global spectral shape with one broad peak. Due to the global envelope's shape, sudden jumps of a certain size are often perceived as ambiguous in terms of the direction of change. A similar phenomenon occurs with testing pitch perception when using Shepard tones: changes of half an octave are perceived as increasing in pitch by some listeners and decreasing in pitch by others (Chambers et al. 2017). This ambiguity can be resolved by presenting a context prior to the shift from either the lower or higher half octave around the test stimuli. Judgements of shift direction were generally in the region of the prior context, demonstrating a context sensitivity of timbral shift similar to that found for pitch.

2.4 Spectromorphological Conceptions of Timbre

An alternative approach to the conception of timbre as a set of orthogonal perceptual dimensions is to consider it as a complex spectrotemporal representation taken as a whole. Different conceptions of this kind will be considered briefly here as they relate to the notion of perceptual representation (for more detail, refer to Elhilali, Chap. 12).

2.4.1 The Auditory Image Model

The peripheral auditory processing model by Patterson et al. (1995) computes an *auditory image* from an input signal. It comprises stages of: (1) outer and middle ear filtering; (2) spectral analysis with dynamic, compressive, gammachirp filtering to reflect biomechanical processing of the basilar membrane; (3) neural encoding of filtered waves to create a neural activity pattern (NAP) that represents the distribution of activity in the auditory nerve; and (4) strobed temporal integration to compute the time intervals between peaks in the NAP and the creation of time-interval histograms in each filter that form the simulated auditory image (SAI) (Fig. 2.7). In Patterson's (2000) conception, pitch would be represented by the repeating forms (see the peaks in the time-interval histograms in Fig. 2.7) and timbre would be represented by the shape of the form (see frequency-channel histograms in Fig. 2.7). This representation doesn't seem to have been exploited much in timbre research to date, but it potentially captures the representation of acoustic scale discussed in Sect. 2.5.1 (van Dinther and Patterson 2006).

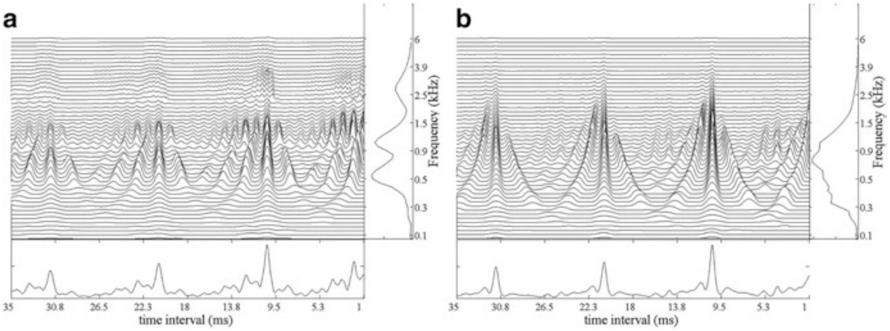

Fig. 2.7 Simulated auditory images of sustained parts of tones produced by a baritone voice **(a)** and a French horn **(b)** at the same fundamental frequency. Each line in the auditory image shows the simulated activity in a given frequency channel (auditory filters) over time. The *lower diagrams* in (a) and (b) represent the global time interval histogram across frequency channels, and the *diagrams to the right* in (a) and (b) represent the global level in each frequency channel (Reproduced from figure 5 in van Dinther and Patterson 2006; used with permission of The Acoustical Society of America)

2.4.2 Multiresolution Spectrotemporal Models

A new class of modulation representations describes sound signals according to their frequency and amplitude variation over time (as in a spectrogram or cochleogram) but also includes a higher-dimensional topography of spectral and temporal modulations, termed *scale* and *rate*, respectively. These representations include the modulation power spectrum (MPS) (Elliott et al. 2013) or simulations of cortical spectrotemporal receptive fields (STRF) (Shamma 2001). The MPS is obtained by computing the amplitude spectrum of the two-dimensional Fourier transform of a time-frequency representation of the sound pressure waveform. The STRF is meant to model the response patterns of primary auditory cortical neurons that are selectively sensitive to particular temporal and spectral modulations. The rate dimension represents temporal modulations derived from the cochlear filter envelopes, and the scale dimension represents modulations present in the spectral shape derived from the spectral envelope (for more detail, see Elhilali, Chap. 12). It has been proposed that models of timbre might be derived from these representations.

Elliott et al. (2013) note that spectral and temporal descriptors are often treated separately in attempts to characterize timbre, but the MPS might be able to characterize sounds physically by integrating these diverse features. They conducted a timbre dissimilarity study on a larger corpus of sustained orchestral instrument sounds than had been attempted before (42 compared to the 12–21 used previously) and decided on a five-dimensional space, claiming that the five-dimensional solution is “necessary and sufficient to describe the perceptual timbre space of sustained orchestral tones” (Elliott et al. 2013, p. 389). Several notes of caution are warranted, however, with regard to the necessity and sufficiency of this five-dimensional space. First, a three-dimensional solution explained 91.5% of the squared distance between instruments, so the two higher dimensions were making small contributions. Second, these sounds were all at a single midrange pitch (making it in a very high pitch register of some low instruments and a very low register of some high instruments) and presumably at a given dynamic marking, so things might change at different pitches and dynamic markings. Lastly, it is highly likely that, based on Lakatos' (2000) results, different dimensions would have to be added if percussion instruments were added to the set or if impulsive sounds were produced on these same instruments, such as string pizzicati.

Elliott et al. (2013) computed the MPS for each of their 42 sounds and for more traditional audio descriptors such as statistical moments of the spectrum and the temporal envelope, attack time, and spectral and temporal entropy. Many features, such as the harmonicity of the signals and spectral shape, show up as specific scale characteristics in the MPS. Temporal features, such as vibrato (frequency modulation), tremolo (amplitude modulation), and the shape of the temporal envelope, show up as rate characteristics. Twenty principal components (PC) derived from the MPSs were selected for regression analysis onto the five dimensions of the timbre space. Significant regressions of the PCs were obtained for all dimensions but the third. Subsequent regressions of traditional audio descriptors (see Sect. 2.3.3; Caetano, Saitis, and Siedenburg, Chap. 11) on the five perceptual dimensions were significant for all dimensions except the fifth. Elliott et al. (2013) concluded that the MPS and audio descriptor analyses are complementary, but certain properties of the timbre spaces are clearer with the MPS representations. It is notable, however, that the explanatory power of the two approaches is roughly equivalent. This leaves open the question of whether timbre indeed emerges from a high-dimensional spectrotemporal form or whether it is a limited set of orthogonal perceptual dimensions.

Patil et al. (2012) used a combination of STRF modeling and machine learning to model timbre dissimilarity data. They presented listeners with pairs of eleven musical-instrument sounds at each of three pitches. They combined the data across pitches and across listeners for the modeling analysis. With a machine-learning algorithm, they derived a confusion matrix among instruments based on instrument distances in the STRF representation. This matrix was then compared to the dissimilarity data. The STRF model achieved a very strong correlation with the human data. However, the predictions of timbre dissimilarity ratings relied heavily on dimensionality-reduction techniques driven by the machine-learning algorithm. For example, a 3840-dimensional representation with 64 frequency filters, 10 rate filters, and 6 scale filters was projected into a 420-dimensional space, essentially yielding a result that is difficult to interpret from a psychological standpoint. It remains to be determined to what extent this approach can be generalized to other timbre spaces (although for applications to instrument recognition, see Agus, Suied, and Pressnitzer, Chap. 3).

2.5 Sound Source Perception

A growing literature documents the ability of untrained listeners to recognize a variety of mechanical properties of sound sources. The development of a theory of sound source perception thus concerns what relevant acoustic information is created by setting sounding objects into vibration and what principles govern the mapping from acoustic information to perceptual response. The perceptual process requires at least two decisions: Which acoustic properties are to be taken into account, and how acoustic information should be weighted perceptually for a given use of that information (e.g., comparing qualities, identifying materials, or size of the object)? These decision-making processes are acquired and refined as a result of one's interactions with the environment.

According to the *information processing* approach to psychology, the link between the perceptual qualities of a sound source, its abstract representation in memory, its identity, and the various meanings or associations it has with other objects in the listener's environment are hypothesized to result from a multistage process (McAdams 1993). This process progressively analyzes and transforms the sensory information initially encoded in the auditory nerve. Perception arises from the extraction of relevant features of the sound in the auditory brain, and recognition is accomplished by matching this processed sensory information with some representation stored in a lexicon of sound forms in long-term memory.

Another approach is that of *ecological psychology* (Gaver 1993). Ecological theory hypothesizes that the physical nature of the sounding object, the means by which it has been set into vibration, and the function it serves for the listener are perceived directly, without any intermediate processing. In this view, perception does not consist of an analysis of the elements composing the sound event followed by their subsequent reconstitution into a mental image that is compared with a representation in memory. Ecological psychologists hypothesize that the perceptual system is tuned to those aspects of the environment that are of biological significance to the organism or that have acquired behavioral significance through experience. However, the claim that the recognition of the function of an object in the environment is perceived directly without processing seems to evacuate the whole question of *how* organisms with auditory systems stimulated by sound vibrations come to be aware of the significance of a sound source or how such sources acquire significance for these listeners. Ecological acoustics places more emphasis on the mechanical structure of sound-producing objects and the acoustic events they produce, which are relevant to a perceiving (and exploring) organism (Carello et al. 2005).

A middle ground between these two approaches is what might be termed *psychomechanics* (McAdams et al. 2004). The aim is to establish quantitative relations between the mechanical properties of sound sources and their perceptual properties, recognizing that listeners most often attend to vibrating objects rather than the sound properties themselves (although the latter clearly play a strong role in music listening) (Gaver 1993). The link between mechanics and acoustics is deterministic, and so there is a very tight relation between mechanics, acoustics, and auditory perception.

Timbral properties, together with those related to pitch, loudness, and duration, contribute to the perception and identity of sound sources and the actions that set them into vibration. In this chapter, the focus is on perception of the properties that are determined by the geometry and materials of sound sources and the manner in which they are made to vibrate. Agus, Suied, and Pressnitzer (Chap. 3) provide more detail on timbre categorization and recognition.

2.5.1 Sound Source Geometry

There are many geometric properties of sound sources to which listeners are sensitive, including shape and size. Repp (1987) demonstrated that under certain conditions listeners can judge hand configuration from the sound of two hands clapping together. This ability is based on the spectral distribution of the hand clap: more cupped hands produce lower resonances than less cupped hands or fingers on the palm.

Listeners are also sensitive to differences in the width and thickness of rectangular metal and wood bars of constant length (Lakatos et al. 1997). The relevant information used to decide which visual depiction of two bars of differing geometry corresponds to that of two sounds presented in sequence was related to the different modes of vibration of the bars; but audiovisual matching performance is better for more homogeneous (isotropic) materials, such as steel, than with anisotropic materials, such as grainy soft woods. This latter finding can be explained by the more reliable modal information provided by isotropic materials.

Cabe and Pittenger (2000) studied listeners' perceptions of the filling of cylindrical vessels using changes in geometry to estimate by sound when a vessel would be full (presumably related to the resonant frequency of the tube above the water level). Listeners had to distinguish different events generated by pouring water into an open tube. Categorization accuracy of whether the sound indicated filling, emptying, or a constant level ranged from 65% to 87%, depending on the type of event. When listeners were asked to fill the vessel up to the brim using only auditory information, filling levels were close to the maximum possible level, suggesting they could hear when the vessel was full. If blind and blindfolded subjects were asked to fill to the brim vessels of different sizes and with different water flow velocities, again overall performance was accurate, and no significant differences between blind and blindfolded participants were found.

Kunkler-Peck and Turvey (2000) investigated shape recognition from impact sounds generated by striking steel plates of constant area and variable height/width with a steel pendulum. Listeners had to estimate the dimensions of the plates. Their performance indicated a definite impression of the height and width of plates.

Judgements of the dimensions of plates were modulated by the type of material (steel, Plexiglas, wood) but maintained the height/width ratio, that is, the relative shape. Performance in both of these tasks was predicted by the frequencies of the vibration modes of the plates. Additional experiments addressed shape recognition directly. For stimuli generated by striking triangular, circular, or rectangular steel plates of constant area, shape was correctly classified above chance level. With stimuli produced by striking the same shapes of plates made of steel, wood, and Plexiglas, the material was almost perfectly classified, and shape was correctly classified above chance level, demonstrating that material recognition is more robust than shape recognition.

Another important aspect of geometry is the size of a sound source. There are acoustic properties that communicate size information in natural sounds involving forced-vibration systems such as human and animal vocalizations and wind and bowed-string musical instruments. As animals grow, their vocal tracts increase in length. In the case of humans, for example, this increase is accompanied by predictable decreases in the formant frequencies of speech and sung sounds (see Mathias and von Kriegstein, Chap. 7). Smith et al. (2005) used a vocoder-based technique (*STRAIGHT*) (Kawahara et al. 1999) to manipulate acoustic scale in vowel sounds, even well beyond the range of sizes normally encountered in humans. Acoustic scale, in their conception, has two components: the scale of the excitation source (pulse rate decreases as source size increases) and the scale of the resonant filter (resonant frequency decreases with size). They showed that listeners not only reliably discriminate changes in acoustic scale associated with changes in vocal tract length but can still recognize the vowels in the extreme low and high ranges of the acoustic scale. This finding suggests an auditory ability to normalize glottal pulse rate (related to pitch) and resonance scale (related to timbre). Van Dinther and Patterson (2006) found a similar relation between acoustic scale and size perception for musical sounds. Listeners can reliably discriminate acoustic scale for musical sounds, although not as well as they can discriminate acoustic scale for vocal sounds. In addition, they can still identify instruments whose sounds have been transformed digitally in acoustic scale beyond the range of normal instruments.

Along the same lines, Plazak and McAdams (2017) found that listeners are sensitive to change in size of a given instrument (created with a version of the *STRAIGHT* algorithm), but that this depends on the instrument (better for oboe and voice with formant structures—resonance peaks in the spectral envelope—than for French horn, cello, and alto saxophone with more low-pass spectral shapes). It is worth mentioning that the notion of “size” has been employed as a concept in an orchestration treatise by Koechlin (1954), as *volume* in French or *extensity* in English, and has been linked to spectral shape in both ordinal and ratio scaling experiments (Chiasson et al. 2017). It would be interesting to test this hypothesis in timbre space studies, including similar instruments of various sizes, created either mechanically or with digital means such as the *STRAIGHT* algorithm.

2.5.2 Sound Source Materials

Sound can convey information about the materials composing an object that are often not directly available to the visual system. Several experiments have explored listeners' perceptions of material properties of struck objects (plates and bars). Mechanical factors that determine material properties include but are not limited to: (1) *modal frequencies* that depend on wave velocity (related to the elasticity and mass density of the material), although these frequencies can also vary with geometry; and (2) the way that *damping* (energy loss due to internal friction) varies with the modal frequencies.

In one of the first studies to address perception of mechanical properties, Freed (1990) measured the attack-related timbral dimension of mallet hardness. Stimuli were generated by striking four metal cooking pots of various diameters with six mallets of variable hardness. Hardness ratings corresponded to relative mallet hardness and were found to be independent of the pan size, thus revealing the subjects' ability to judge the material properties of the mallet independently of those of the sounding object. Hardness increases with the global spectral level and the spectral centroid (both averaged over the first 325 ms of the signal) and decreases with the slope of the change in spectral level over time and the temporal centroid of the time-varying spectral centroid (the centroid-weighted average time). Harder mallets are more intense, have higher spectral centroids, sharper decreasing spectral level slopes, and earlier temporal centroids.

Sound sources are perceived by integrating information from multiple acoustic features. Thus, part of the task of understanding the integration of information becomes that of unraveling the principles that govern the assignment of perceptual weights to sound properties. Two factors have a potential influence on this process: (1) the accuracy of the acoustic information within the environment in which the perceptual criteria develop and (2) the ability of a perceptual system to exploit the acoustic information. Information accuracy is the extent to which levels of a source property are reliably diversified by levels of a sound property within the learning environment. For example, if the task is to rate the hardness of an object, information accuracy can be given by the absolute value of the correlation between values of the physical hardness and values of a specific acoustic feature. Based on previous hypotheses concerning the perceptual weight of accurate information, one might expect that a listener would weight acoustic information in proportion to its accuracy. For example, if frequency specifies the size of an object twice as accurately as sound level, perceptual estimation of size would weight frequency twice as heavily as level.

Another factor potentially influencing the structure of perceptual criteria is the ability to exploit the information carried by different acoustic features. This factor can be determined from a listener's ability to discriminate a source property and to benefit from training in such a task. One might expect that, independently of the task at hand, a listener would weight more heavily the acoustic information that is more easily exploited. The factors that influence the integration of acoustic information are largely unknown.

Giordano et al. (2010) investigated the extent to which the perceptual weighting of acoustic information is modulated by its accuracy and exploitability. They measured how the perceptual weighting of different features varies with the accuracy of information and with a listener's ability to exploit that information. Participants judged the hardness of a hammer and a sounding object whose interaction generates an impact sound. In the first experiment in which trained listeners were asked to discriminate hammer or object hardness, listeners focused on the most accurate information, although they had greater difficulty when discriminating hammer hardness. The authors inferred a limited exploitability for the most accurate hammer-hardness information. In a subsequent hardness rating experiment, listeners focused on the most accurate information only when estimating object hardness. In an additional hardness rating experiment, sounds were synthesized by independently manipulating source properties that covaried in the previous two experiments: object hardness and impact properties, such as contact time of the hammer with the object and the extent to which the hammer is compressed during the impact at a given striking force (the *force stiffness coefficient*). Object hardness perception relied on the most accurate acoustic information, whereas impact properties more strongly influenced the perception of hammer hardness. Overall, perceptual weight increased with the accuracy of acoustic information, although information that was not easily exploited was perceptually secondary, even if accurate.

Klatzky et al. (2000) investigated material similarity perception using synthesized stimuli composed of a series of exponentially damped sinusoids with variable frequency and frequency-dependent decay of the constituent partials that were designed to mimic impacted plates of different materials. The frequency-dependent decay is related to damping and depends exclusively on material, being relatively independent of geometry. Listeners rated the perceived difference in the materials of two sounds. An MDS analysis revealed dimensions corresponding to the two synthesis parameters. The results did not differ significantly between experiments in which the sounds were either equalized in overall energy or were not equalized, leading to the conclusion that intensity is not relevant in the judgment of material difference.

In further experiments by Klatzky and colleagues, listeners rated the difference in the perceived length of the objects and categorized the material of the objects using four response alternatives: rubber, wood, glass, and steel. Results indicated that ratings of material difference and length difference were significantly influenced by both damping and frequency, even though the contribution of the decay parameter to ratings of length difference was smaller than to ratings of material difference. An effect of both of these variables was found in the categorization task. Lower decay factors led to more steel and glass identifications compared to those for rubber and wood, whereas glass and wood were chosen for higher frequencies than were steel and rubber. Therefore, both factors are necessary to specify these material categories.

Material and geometric properties of synthesized impacted bars with a tube resonator (as with a xylophone or marimba) were varied by McAdams et al. (2004). They inferred the perceptual structure of a set of sounds from an MDS analysis of dissimilarity ratings and quantified the psychomechanical relations between sound source properties and perceptual structure. Constant cross-section bars that varied in mass density and the viscoelastic damping coefficient were synthesized with a physical model in one experiment. A two-dimensional perceptual space resulted, and the dimensions were correlated with the mechanical parameters after applying a power-law transformation. Variable cross-section bars (as in a xylophone bar) varying in length and viscoelastic damping coefficient were synthesized in another experiment with two sets of lengths creating high- and low-pitched bars. With the low-pitched bars, there was a coupling between the bar and the resonator that modified the decay characteristics. Perceptual dimensions again corresponded to the mechanical parameters. A set of potential temporal, spectral, and spectrotemporal descriptors of the auditory representation were derived from the signal. The dimensions related to both mass density and bar length were correlated with the frequency of the lowest partial and were related to pitch perception. The descriptor most likely to represent the viscoelastic damping coefficient across all three stimulus sets was a linear combination of a decay constant derived from the temporal envelope and the spectral center of gravity derived from a cochlear filterbank representation of the signal.

McAdams et al. (2010) synthesized stimuli with a computer model of impacted plates in which the material properties could be varied. They manipulated viscoelastic and thermoelastic damping and wave velocity. The range of damping properties represented an interpolated continuum between materials with predominant viscoelastic and thermoelastic damping (glass and aluminum, respectively). The perceptual structure of the sounds was inferred from an MDS analysis of dissimilarity ratings and from their categorization as glass or aluminum. Dissimilarity ratings revealed dimensions that were closely related to mechanical properties: a wave-velocity-related dimension associated with pitch and a damping-related dimension associated with timbre and duration (Fig. 2.8). When asked to categorize sounds according to material, however, listeners ignored the cues related to wave velocity and focused on cues related to damping (Fig. 2.9). In both dissimilarity rating and identification experiments, the results were independent of the material of the mallet striking the plate (rubber or wood). Listeners thus appear to select acoustic information that is reliable for a given perceptual task. Because the frequency changes responsible for detecting changes in wave velocity can also be due to changes in geometry, they are not as reliable for material identification as are damping cues. These results attest to the perceptual salience of energy loss phenomena in sound source behavior.

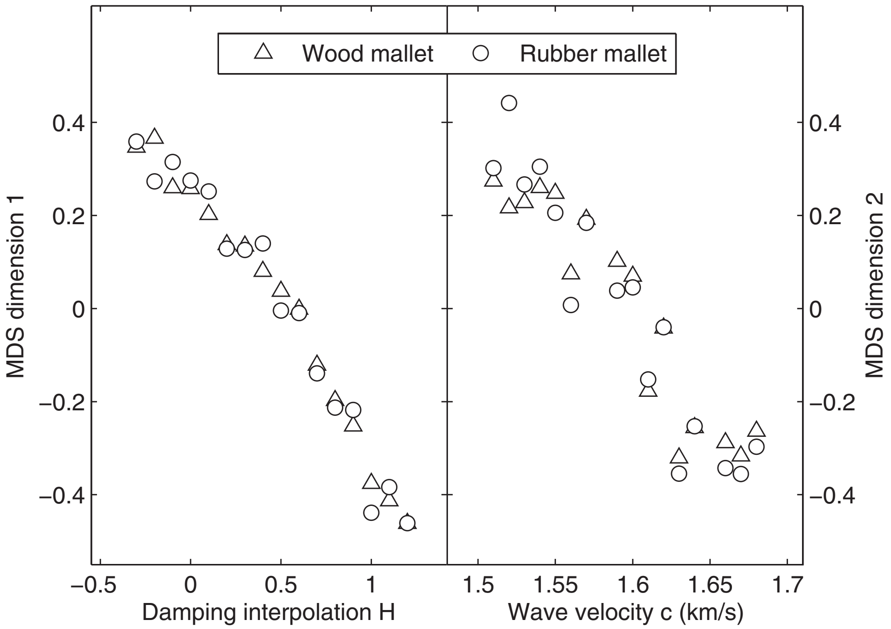

Fig. 2.8 Relation between MDS dimensions and physical model parameters. In the **left panel**, the coordinate along Dimension 1 is plotted as a function of the factor that controlled the interpolation between viscoelastic and thermoelastic damping (*damping interpolation H*). In the **right panel**, the coordinate along Dimension 2 is plotted as a function of wave velocity (km/sec). Note the nearly linear relationship between the two variables in both cases. Note also that striking the object with a wood or rubber mallet did not influence the perceived dissimilarities in each set of sounds (Reproduced from figure 5 in McAdams et al. 2010; used with permission of The Acoustical Society of America)

2.5.3 Actions on Vibrating Objects

Although most perceptual studies of mechanical properties of sounding objects have focused primarily on material and geometric properties of the objects themselves, some research has addressed the actions by which the objects are set into vibration, such as scraping, rolling, hitting, and bouncing (for more on simulation of these phenomena, see Ystad, Aramaki, and Kronland-Martinet, Chap. 13). In everyday life, we are more likely to listen to the properties of sources that generate sound than to the properties of the sound itself. So the question becomes: To what *properties of actions* that excite sounding objects are listeners sensitive and which *sound properties* carry the relevant information for those actions?

Stoelinga et al. (2003) measured the sounds of metal balls rolling over fiberboard plates. A spectrographic analysis of the resulting sounds revealed time-varying ripples in the frequency spectrum that were more closely spaced when the ball was in the middle of the plate than when it was closer to the edge. The ripple spacing was

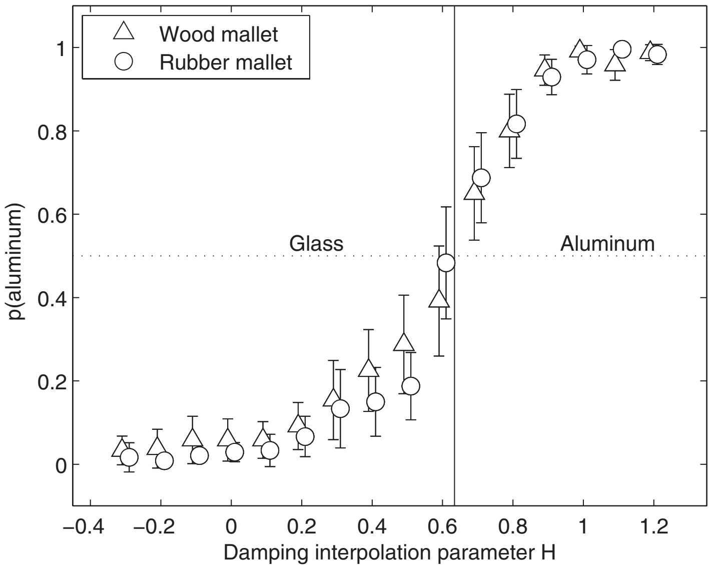

Fig. 2.9 Probability of classifying a sound as aluminum. Error bars are 95% confidence intervals. The probability (p) that a sound was classified as aluminum is plotted as a function of the damping interpolation parameter (H). The *vertical line* indicates the point of inflection of the curve and thus the category boundary between glass and aluminum. There is no significant difference in the functions for wood versus rubber mallets (Reproduced from figure 7 in McAdams et al. 2010; used with permission of The Acoustical Society of America)

also tighter for lower frequencies than for higher frequencies. This pattern arises from the interference between the sound directly generated at the point of contact between the ball and plate and first-order reflections of the sound at the edge of the plate. The authors hypothesized that this effect is a crucial cue in the synthesis of realistic rolling sounds.

Addressing this issue from a perceptual perspective, Houben et al. (2004) conducted three experiments on the auditory perception of the size and speed of wooden balls rolling over a wooden plate. They recorded balls of various sizes rolling at different speeds. One experiment showed that when pairs of sounds are presented, listeners are able to choose the one corresponding to the larger ball. A second experiment demonstrated that listeners can discriminate between the sounds of balls rolling at different speeds, although some listeners had a tendency to reverse the labeling of the speed. The interaction between size and speed was tested in a final experiment in which the authors found that if both the size and the speed of a rolling ball are varied, listeners generally are able to identify the larger ball, but the judgment of speed is influenced by the size. They subsequently analyzed the spectral and temporal properties of the recorded sounds to determine the cues available to listeners to make their judgements. In line with the observed interaction effect, the results suggested a conflict in available cues when varying both size and speed. The authors were able to rule out auditory roughness as a cue because the acoustic differences that would affect roughness perception were smaller than the just noticeable difference for roughness predicted by Zwicker and Fastl (1990). So it is unlikely that this auditory attribute is responsible for the interaction. However, the spectral shape of the rolling sounds is affected by both speed and size of the rolling balls with greater emphasis of higher frequencies for smaller diameters and faster speeds. The spectral differences were apparently greater than the discrimination threshold, making this a likely candidate for the interaction.

Lemaitre and Heller (2012) addressed the issue of the relative importance of actions that generate sounds and the properties of the sounding objects. They conducted a study that compared the performance of listeners who were asked to identify either the actions or the materials used to generate sound stimuli. Stimuli were recorded from a set of cylinders with two sizes and four materials (wood, plastic, glass, metal). Each object was subjected to four different actions (scraping, rolling, hitting, bouncing). The authors reported that listeners were faster and more accurate at identifying the actions than the materials, even if they were presented with a subset of sounds for which both actions and materials were identified at similarly high levels. They concluded that the auditory system is well suited to extract information about sound-generating actions.

In a subsequent study, Lemaitre and Heller (2013) examined whether the auditory organization of categories of sounds produced by actions includes a privileged or *basic* level of description. They employed sound events consisting of materials (solids, liquids, gases) undergoing simple actions (friction, deformation, impacts for solids; splashing, dripping or pouring liquids; whooshing, blowing, puffing or exploding gases). Performance was measured either by correct identification of a sound as belonging to a category or by the extent to which it created lexical priming. The categorization experiment measured the accuracy and reaction time to brief excerpts of the sounds. The lexical priming experiment measured reaction time benefits and costs caused by the presentation of these sounds immediately prior to a lexical decision (whether a string of letters formed a word or not). The level of description of a sound was varied in terms of how specifically it described the physical properties of the action producing the sound (related or unrelated sounds and words). Listeners were better at identification and showed stronger priming effects when a label described the specific interaction causing the sound (e.g., gushing or tapping) in comparison either to more general descriptions (e.g. pour, liquid, where gushing is a specific way of pouring liquid; or impact, solid, where tapping is a way of impacting a solid) or to more detailed descriptions that employed adverbs regarding the manner of the action (e.g., gushing forcefully or tapping once). These results suggest a quite robust and complex encoding of sound-producing actions at both perceptual and semantic levels.

The application of the psychomechanical approach has focused on fairly simple sounding objects and actions, in many cases specifically targeting sound events that can be synthesized with physical models such as impacted bars and plates. Future work of this nature on more complex systems, such as musical instruments, will require more refined models of these complex sound sources, particularly with regard to changes in timbral properties that covary with other parameters such as fundamental frequency and playing effort.

2.6 Interaction with Other Auditory Attributes

Most studies of musical timbre have constrained pitch and loudness to single values for all of the instrument sounds with the aim of focusing listeners' attention on timbre alone, which is the legacy of the negative definition of timbre as what's left over when these parameters are equalized. This raises an important question, however: Do the timbral relations revealed for a single pitch and/or a single dynamic level (related to playing effort) hold at different pitches and dynamic levels? And more importantly, if one intended to extend this work to real musical contexts, would the relations hold for timbres being compared across pitches and dynamic levels, particularly given the fact that timbre covaries with both pitch and dynamics in musical instruments? A subsidiary issue would be to determine what spectral, temporal, and spectrotemporal properties of the sounds covary with these other musical parameters. The multiple interactions of timbre, pitch, and loudness have been demonstrated with a *speeded classification paradigm* by Melara and Marks (1990). They found that having random or correlated variation in a second dimension affected speed and accuracy of classification along a primary, criterial dimension for pairs of these auditory parameters.

2.6.1 Timbre and Pitch

Some timbre dissimilarity studies have included sounds from different instruments at several pitches. Marozeau et al. (2003) demonstrated that timbre spaces for recorded musical-instrument tones are similar at three different pitches (B3, C#4, Bb4, where C4 is middle C). Listeners were also able to ignore pitch differences within an octave when they were asked to compare only the timbres of the tones: B3 to Bb4 is a major 7th, one semitone short of an octave. However, when the pitch variation is greater than an octave, interactions between the two attributes occur. Marozeau and de Cheveigné (2007) varied the spectral centroid of a set of synthesized sounds while also varying the fundamental frequency over a range of eighteen semitones (an octave and a half). Pitch appears in the MDS space as a dimension orthogonal to the timbre dimension, which indicates that listeners were not able to ignore the pitch change but treated it more or less orthogonally to timbre. Paradoxically, however, pitch differences were found to systematically affect the timbre dimension related to the spectral centroid with slight shifts toward lower perceptual values along this dimension for higher pitches (Fig. 2.10). This result perhaps suggests that listeners, who were instructed to ignore pitch and focus on timbre, had a tendency to compensate for the change in brightness induced by the higher pitches in their dissimilarity ratings or that this dimension is related to the richness of the spectrum with sounds at higher pitches having more sparse spectra. Handel and Erickson (2001) had also found that nonmusician listeners had difficulty extrapolating the timbre of a sound source across large differences in pitch in a

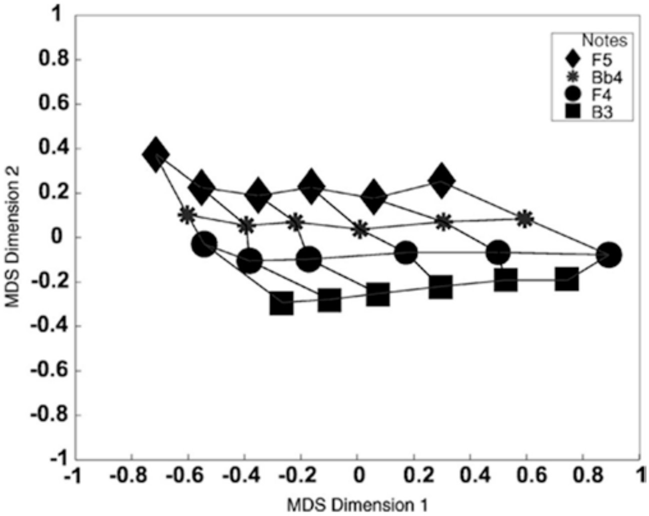

Fig. 2.10 Multidimensional scaling (MDS) solution in two dimensions rotated to maximize correlation between Dimension 1 and the spectral centroid. Note that the musical notes at different fundamental frequencies (different symbols) are not strongly affected by spectral centroid: the curves are flat. Note also that they are clearly separated from each other along MDS dimension 2, indicating a relative independence of pitch and brightness. The different spectral centroid values at each fundamental frequency behave very regularly and are fairly evenly spaced (along MDS dimension 1), but there is an increasing shift to lower perceptual values as the fundamental frequency increases. Frequencies of Notes: *B3*, 247 Hz; *Bb4*, 349 Hz; *F4*, 466 Hz; *F5*, 698 Hz (Reproduced from figure 3 in Marozeau and de Cheveigné 2007; used with permission of The Acoustical Society of America)

recognition task, although Steele and Williams (2006) found that musician listeners could extrapolate timbre with intervals of more than two octaves. Therefore, there are limits to timbral invariance across pitch, but they depend on musical training.

Inversely, timbre can also affect pitch perception. Vurma et al. (2011) reported that timbre differences on two successive tones can affect judgements of whether two pitches are in tune. When the second tone in a pair with identical fundamental frequencies had a brighter timbre than the first, it was judged as sharp (higher pitch) and for the inverse case, it was judged as flat (lower pitch). This result confirmed an effect reported by Russo and Thompson (2005) in which ratings of interval size by nonmusicians for tones of different pitches were greater when the timbral brightness changed in the same direction and were diminished when brightness change was incongruent.

Finally, some studies have demonstrated mutual interference of pitch and timbre. Krumhansl and Iverson (1992) found that uncorrelated variation along pitch or timbre symmetrically affected speeded classification of the other parameter. Allen and Oxenham (2014) obtained similar results when measuring difference limens in stimuli that had concurrent random variations along the unattended dimension. These authors found symmetric mutual interference of pitch and timbre in the discrimination task when making sure that changes in timbre and pitch were of similar perceptual magnitude. Their results suggest a close relation between timbral brightness and pitch height (for more on the semantics of brightness, see Saitis and Weinzierl, Chap. 5). This link would be consistent with underlying neural representations for pitch and timbre that share common attributes such as the organization of tonotopy and periodicity in the brain. Such a shared neural representation might underlie the perception of *register* (in which octave a particular pitch class is being played) (Robinson 1993; Patterson et al. 2010).

2.6.2 Timbre and Playing Effort (Dynamics)

Changes in dynamics can also produce changes in timbre for a given instrument. Sounds produced with greater playing effort (e.g., fortissimo versus pianissimo) have greater energy at all the frequencies present in the softer sound, but the spectrum also spreads toward higher frequencies as more vibration modes of the physical system are excited. This mechanical process creates changes in several descriptors of spectral shape, including a higher spectral centroid, greater spectral spread, and a lower spectral slope. There do not appear to be studies that have examined the effect of change in dynamic level on timbre perception, but some work has studied the role of timbre in the perception of dynamic level independently of the physical level of the signal.

Fabiani and Friberg (2011) varied pitch, sound level, and instrumental timbre (clarinet, flute, piano, trumpet, violin) and studied the effect of these parameters on the perception of the dynamics of isolated instrumental tones. Listeners were asked to indicate the perceived dynamics of each stimulus on a scale from pianissimo (*pp*) to fortissimo (*ff*). The timbral effects produced at different dynamics, as well as the physical level, had equally large effects for all five instruments, whereas pitch was relevant mostly for clarinet, flute, and piano. Higher pitches received higher dynamic ratings for these three instruments. Thus, estimates of the dynamics of musical tones are based both on loudness and timbre and, to a lesser degree, on pitch as well.

2.7 Summary and Conclusions

Timbre is clearly a complex phenomenon that is multidimensional, including many different aspects such as brightness, attack quality, hollowness, and even aspects of the size, shape, and material composition of sound sources. Studies of timbre discrimination reveal listeners' heightened sensitivity to subtle spectral and temporal properties of musical-instrument sounds. However, in musical contexts, the sensitivity to temporal envelope details seems to be diminished. One approach to timbre's inherent multidimensionality is to use MDS of dissimilarity ratings to model perceptual relations in terms of shared dimensions, specific features and weights on the dimensions, and features for different individuals or groups of individuals.

Common dimensions have been associated with various audio descriptors through correlation analyses, with more or less success depending on the sound set used. Audio descriptors, such as the spectral centroid, attack and/or decay time, and deviation from a smooth spectral envelope, seem ubiquitous for many classes of musical-instrument sounds and have been validated by confirmatory studies. However, some caution is warranted in the audio descriptor realm: the plethora of descriptors in the literature do not all vary independently even across a very large database of musical sounds at various pitches and dynamic levels, and there may be only about ten independent classes of such descriptors (for musical instrument sounds at least). Furthermore, at this point only scalar values of such descriptors have been employed, and new research needs to examine the time-varying properties of natural sounds, which carry much information concerning the state of sounding objects. In some cases, common dimensions have also been associated with the mechanical properties of sound sources, such as damping rate for material properties, relations among modal frequencies of solids or resonance frequencies of air columns for geometric properties, and temporal and textural properties of the actions that set objects into vibration. Indeed, in some cases it appears that listeners are more sensitive to what is happening to objects in the environment (actions) than to the nature of the objects themselves.

In examining the extent to which modeled representations depend on stimulus context, it seems that timbre dissimilarity ratings, in particular, are fairly robust to the range of sounds present. This result may suggest that there are aspects of timbre perception that are absolute and tied to recognition and categorization of sound sources through interactions of perception with long-term memory accumulated through experiences with those sources. However, timbre relations can be affected by changes along other dimensions such as pitch and loudness. These interactions may be partly due to the sharing of underlying neural representations and partly due to the fact that all of these auditory attributes covary significantly in the sound sources encountered in everyday life and in music listening.

Another class of models presumes that timbre is a complex, but unitary, multidimensional structure that can be modeled with techniques such as auditory images, modulation power spectra, or spectrotemporal receptive fields. This work is still in its infancy, and it is not yet clear what new understanding will be brought to the realm of timbre by their use or whether alternative models will provide more explanatory power than the more traditional multidimensional approach.

Acknowledgements The writing of this chapter was supported by grants from the Canadian Natural Sciences and Engineering Research Council (RGPIN-2015-05280 and RGPAS-478121-15), the Canada Research Chairs Program (#950-223484), and the Killam Research Fellow Program of the Canada Council for the Arts.

Compliance with Ethics Requirements Stephen McAdams declares that he has no conflict of interest.

References

Allen EJ, Oxenham AJ (2014) Symmetric interactions and interference between pitch and timbre. J Acoust Soc Am 135(3):1371–1379. https://doi.org/10.1121/1.4863269

Almeida A, Schubert E, Smith J, Wolfe J (2017) Brightness scaling of periodic tones. Atten Percept Psychophys 79(7):1892–1896. https://doi.org/10.3758/s13414-017-1394-6

Cabe PA, Pittenger JB (2000) Human sensitivity to acoustic information from vessel filling. J Exp Psychol Hum Percept Perform 26(1):313–324. https://doi.org/10.1037//0096-1523.26.1.313

Caclin A, McAdams S, Smith B, Winsberg S (2005) Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones. J Acoust Soc Am 118(1):471–482. https://doi.org/10.1121/1.1929229

Caclin A, Brattico E, Ternaviemi M, Näätänen R, Morlet D, Giard MH, McAdams S (2006) Separate neural processing of timbre dimensions in auditory sensory memory. J Cognitive Neurosci 18(12):1959–1972. https://doi.org/10.1162/jocn.2006.18.12.1959

Caclin A, Giard MH, Smith B, McAdams S (2007) Interactive processing of timbre dimensions: a Garner interference study. Brain Res 1138(1):159–170. https://doi.org/10.1016/j.brainres.2006.12.065

Carello C, Wagman JB, Turvey MT (2005) Acoustic specification of object properties. In: Anderson JD, Anderson BF (eds) Moving image theory: ecological considerations. Southern Illinois University Press, Carbondale, pp 79–104

Chambers C, Akram S, Adam V, Pelofi C, Sahani M, Shamma S, Pressnitzer D (2017) Prior context in audition informs binding and shapes simple features. Nat Commun 8:15027. https://doi.org/10.1038/ncomms15027

Chiasson F, Traube C, Lagarrigue C, McAdams S (2017) Koechlin's volume: perception of sound extensity among instrument timbres from different families. Music Sci 21(1):113–131. https://doi.org/10.1177/1029864916649638

van Dinther R, Patterson RD (2006) Perception of acoustic scale and size in musical instrument sounds. J Acoust Soc Am 120(4):2158–2176. https://doi.org/10.1121/1.2338295

Elliott TM, Hamilton LS, Theunissen FE (2013) Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. J Acoust Soc Am 133(1):389–404. https://doi.org/10.1121/1.4770244

Esling P, Agon C (2013) Multiobjective time series matching for audio classification and retrieval. IEEE Trans Audio Speech Lang Process 21(10):2057–2072. https://doi.org/10.1109/TASL.2013.2265086

Fabiani M, Friberg A (2011) Influence of pitch, loudness, and timbre on the perception of instrument dynamics. J Acoust Soc Am 130(4):EL193–EL199. https://doi.org/10.1121/1.3633687

Freed DJ (1990) Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. J Acoust Soc Am 87(1):311–322. https://doi.org/10.1121/1.399298

Gaver WW (1993) What in the world do we hear?: an ecological approach to auditory event perception. Ecol Psychol 5(1):1–29. https://doi.org/10.1207/s15326969eco0501_1

Giordano BL, McAdams S (2010) Sound source mechanics and musical timbre perception: evidence from previous studies. Music Percept 28(2):155–168. https://doi.org/10.1525/mp.2010.28.2.155

Giordano BL, Rocchesso D, McAdams S (2010) Integration of acoustical information in the perception of impacted sound sources: the role of information accuracy and exploitability. J Exp Psychol Human 36(2):462–476. https://doi.org/10.1037/A0018388

Grey JM (1977) Multidimensional perceptual scaling of musical timbres. J Acoust Soc Am 61(5):1270–1277. https://doi.org/10.1121/1.381428

Grey JM (1978) Timbre discrimination in musical patterns. J Acoust Soc Am 64(2):467–472. https://doi.org/10.1121/1.382018

Grey JM, Gordon JW (1978) Perceptual effects of spectral modifications on musical timbres. J Acoust Soc Am 63(5):1493–1500. https://doi.org/10.1121/1.381843

Grey JM, Moorer JA (1977) Perceptual evaluations of synthesized musical instrument tones. J Acoust Soc Am 62(2):454–462. https://doi.org/10.1121/1.381508

Hajda JM, Kendall RA, Carterette EC, Harshberger ML (1997) Methodological issues in timbre research. In: Deliège I, Sloboda J (eds) The perception and cognition of music. Psychology Press, Hove, pp 253–306

Handel S, Erickson ML (2001) A rule of thumb: the bandwidth for timbre invariance is one octave. Music Percept 19(1):121–126. https://doi.org/10.1525/mp.2001.19.1.121

Helmholtz HLF von (1885) On the sensations of tone as a physiological basis for the theory of music. Republ.1954 by Dover, New York, from 1877 trans by AJ Ellis from 4th German ed

Houben M, Kohlrausch A, Hermes DJ (2004) Perception of the size and speed of rolling balls by sound. Speech Comm 43:331–345. https://doi.org/10.1016/j.specom.2004.03.004

Kawahara H, Masuda-Katsuse I, de Cheveigné A (1999) Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Comm 27:187–207. https://doi.org/10.1016/S0167-6393(98)00074-0

Kazazis S, Esterer N, Depalle P, McAdams S (2017) A performance evaluation of the Timbre Toolbox and the MIRtoolbox on calibrated test sounds. In: Scavone G, Maestre E, Kemp C, Wang S (eds) Proceedings of the 2017 International Symposium on Musical Acoustics (ISMA). McGill University, Montreal, QC, pp 144–147

Kendall RA (1986) The role of acoustic signal partitions in listener categorization of musical phrases. Music Percept 4(2):185–213. https://doi.org/10.2307/40285360

Kendall RA, Carterette EC (1991) Perceptual scaling of simultaneous wind instrument timbres. Music Percept 8(4):369–404. https://doi.org/10.2307/40285519

Klatzky RL, Pai DK, Krotkov EP (2000) Perception of material from contact sounds. Presence Teleop Virt 9(4):399–410. https://doi.org/10.1162/105474600566907

Koechlin C (1954-1959) Traité de l'orchestration: En quatre volumes [Treatise on orchestration: In four volumes]. M. Eschig, Paris

Krumhansl CL (1989) Why is musical timbre so hard to understand? In: Nielzén S, Olsson O (eds) Structure and perception of electroacoustic sound and music. Excerpta Medica, Amsterdam, pp 43–53

Krumhansl CL, Iverson P (1992) Perceptual interactions between musical pitch and timbre. J Exp Psychol Hum Percept Perform 18(3):739–751. https://doi.org/10.1037/0096-1523.18.3.739

Kruskal JB (1964) Non-metric multidimensional scaling: a numerical method. Psychometrika 29(2):115–129. https://doi.org/10.1007/BF02289694

Kunkler-Peck AJ, Turvey MT (2000) Hearing shape. J Exp Psychol Hum Percept Perform 26(1):279–294. https://doi.org/10.1111/1467-9280.00040

Lartillot O, Toiviainen P (2007) A Matlab toolbox for musical feature extraction from audio. In: Marchand S (ed) Proceedings of the 10th International Conference on digital audio effects (DAFx-07). Université de Bordeaux 1, Bordeaux, France, pp 237–244

Lemaitre G, Heller LM (2012) Auditory perception of material is fragile while action is strikingly robust. J Acoust Soc Am 131(2):1337–1348. https://doi.org/10.1121/1.3675946

Lemaitre G, Heller LM (2013) Evidence for a basic level in a taxonomy of everyday action sounds. Exp Brain Res 226(2):253–264. https://doi.org/10.1007/s00221-013-3430-7

Marozeau J, de Cheveigné A (2007) The effect of fundamental frequency on the brightness dimension of timbre. J Acoust Soc Am 121(1):383–387. https://doi.org/10.1121/1.2384910

Marozeau F, de Cheveigné A, McAdams S, Winsberg S (2003) The dependency of timbre on fundamental frequency. J Acoust Soc Am 114(5):2946–2957. https://doi.org/10.1121/1.1618239

McAdams S (1993) Recognition of sound sources and events. In: McAdams S, Bigand E (eds) Thinking in sound: the cognitive psychology of human audition. Oxford University Press, Oxford, pp 146–198. https://doi.org/10.1093/acprof:oso/9780198522577.003.0006

McAdams S (2013) Musical timbre perception. In: Deutsch D (ed) The psychology of music, 3rd edn. Academic Press, New York, pp 35–67. https://doi.org/10.1016/B978-0-12-381460-9.00002-X

McAdams S (2015) Perception et cognition de la musique [Perception and cognition of music]. Editions J. Vrin, Paris, France

McAdams S, Winsberg S, Donnadieu S, De Soete G, Krimphoff J (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res-Psych Fo 58(3):177–192. https://doi.org/10.1007/Bf00419633

McAdams S, Beauchamp J, Meneguzzi S (1999) Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. J Acoust Soc Am 105(2):882–897. https://doi.org/10.1121/1.426277

McAdams S, Chaigne A, Roussarie V (2004) The psychomechanics of simulated sound sources: material properties of impacted bars. J Acoust Soc Am 115(3):1306–1320. https://doi.org/10.1121/1.1645855

McAdams S, Roussarie V, Chaigne A, Giordano BL (2010) The psychomechanics of simulated sound sources: material properties of impacted thin plates. J Acoust Soc Am 128(3):1401–1413. https://doi.org/10.1121/1.3466867

Melara RD, Marks LE (1990) Interaction among auditory dimensions: timbre, pitch and loudness. Percept Psychophys 48(2):169–178. https://doi.org/10.3758/BF03207084

Nymoen K, Danielsen A, London J (2017) Validating attack phase descriptors obtained by the Timbre Toolbox and MIRtoolbox. In: Proceedings of the 14th sound and music computing conference 2017. Proceedings of the SMC Conferences. Aalto University, Espoo, Finland, pp 214–219

Patil K, Pressnitzer D, Shamma S, Elhilali M (2012) Music in our ears: the biological bases of musical timbre perception. PLoS Comput Biol 8(11):e1002759. https://doi.org/10.1371/journal.pcbi.1002759

Patterson RD (2000) Auditory images: how complex sounds are represented in the auditory system. Journal of the Acoustical Society of Japan 21(4):183–190. https://doi.org/10.1250/ast.21.183

Patterson RD, Allerhand M, Giguère C (1995) Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. J Acoust Soc Am 98(4):1890–1894. https://doi.org/10.1121/1.414456

Patterson RD, Gaudrain E, Walters TC (2010) The perception of family and register in musical tones. In: Jones MR, Fay RR, Popper AN (eds) Music perception. Springer, New York, pp. 13–50. https://doi.org/10.1007/978-1-4419-6114-3_2

Peeters G, Giordano BL, Susini P, Misdariis N, McAdams S (2011) The Timbre Toolbox: extracting audio descriptors from musical signals. J Acoust Soc Am 130(5):2902–2916. https://doi.org/10.1121/1.3642604

Plazak J, McAdams S (2017) Perceiving changes of sound-source size within musical tone pairs. Psychomusicology: Music, Mind, and Brain 27(1):1–13. https://doi.org/10.1037/pmu0000172

Plomp R (1970) Timbre as a multidimensional attribute of complex tones. In: Plomp R, Smoorenburg GF (eds) Frequency analysis and periodicity detection in hearing. Sijthoff, Leiden, pp 397–410

Repp BH (1987) The sound of two hands clapping: an exploratory study. J Acoust Soc Am 81(4):1100–1109. https://doi.org/10.1121/1.394630

Robinson K (1993) Brightness and octave position: are changes in spectral envelope and in tone height perceptually equivalent? Contemp Music Rev 9(1):83–95. https://doi.org/10.1080/07494469300640361

Russo FA, Thompson WF (2005) An interval size illusion: the influence of timbre on the perceived size of melodic intervals. Percept Psychophys 67(4):559–568. https://doi.org/10.3758/BF03193514

Schubert E, Wolfe J (2006) Does timbral brightness scale with frequency and spectral centroid? Acta Acustica united with Acustica 92(5):820–825

Shamma S (2001) On the role of space and time in auditory processing. Trends Cogn Sci 5(8):340–348. https://doi.org/10.1016/S1364-6613(00)01704-6

Shepard RN (1964) Circularity in judgments of relative pitch. J Acoust Soc Am 36(12):2346–2353. https://doi.org/10.1121/1.1919362

Siedenburg K (2018) Timbral Shepard-illusion reveals ambiguity and context sensitivity of brightness perception. J Acoust Soc Am 143(2):EL93–EL98. https://doi.org/10.1121/1.5022983

Siedenburg K, Jones-Mollerup K, McAdams S (2016) Acoustic and categorical dissimilarity of musical timbre: evidence from asymmetries between acoustic and chimeric sounds. Front Psychol 6:1977. https://doi.org/10.3389/fpsyg.2015.01977

Smith DR, Patterson RD, Turner R (2005) The processing and perception of size information in speech sounds. J Acoust Soc Am 117(1):305–318. https://doi.org/10.1121/1.1828637

Steele K, Williams AK (2006) Is the bandwidth for timbre invariance only one octave? Music Percept 23(3):215–220. https://doi.org/10.1525/mp.2006.23.3.215

Stilp CE, Rogers TT, Kluender KR (2010) Rapid efficient coding of correlated complex acoustic properties. Proc Natl Acad Sci 107(50):21914–21919. https://doi.org/10.1073/pnas.1009020107

Stoelinga CNJ, Hermes DJ, Hirschberg A, Houtsma AJM (2003) Temporal aspects of rolling sounds: a smooth ball approaching the edge of a plate. Acta Acustica united with Acustica 89(5):809–817

van Heuven VJJP, van den Broecke MPR (1979) Auditory discrimination of rise and decay times in tone and noise bursts. J Acoust Soc Am 66 (5):1308-1315. https://doi.org/10.1121/1.383551

Vurma A, Raju M, Kuuda A (2011) Does timbre affect pitch? Estimations by musicians and non-musicians. Psychol Music 39(3):291–306. https://doi.org/10.1177/0305735610373602

Wessel DL (1973) Psychoacoustics and music: a report from Michigan State University. PACE: Bulletin of the Computer Arts Society 30:1–2

Zahm JA (1892) Sound and music. AC McClurg and Company, Chicago

Zwicker E, Fastl H (1990) Psychoacoustics: facts and models. Springer-Verlag, Berlin

Chapter 3
Timbre Recognition and Sound Source Identification

Trevor R. Agus, Clara Suied, and Daniel Pressnitzer

Abstract The ability to recognize many sounds in everyday soundscapes is a useful and impressive feature of auditory perception in which timbre likely plays a key role. This chapter discusses what is known of timbre in the context of sound source recognition. It first surveys the methodologies that have been used to characterize a listener's ability to recognize sounds and then examines the types of acoustic cues that could underlie the behavioral findings. In some studies, listeners were directly asked to recognize familiar sounds or versions of them that were truncated, filtered, or distorted by other resynthesis methods that preserved some cues but not others. In other studies, listeners were exposed to novel sounds, and the build-up of cues over time or the learning of new cues was tracked. The evidence currently available raises an interesting debate that can be articulated around two qualitatively different hypotheses: Are sounds recognized through distinctive features unique to each sound category (but of which there would need to be many to cover all recognized categories) or rather, are sounds recognized through a relatively small number of perceptual dimensions in which different sounds have their own recognizable position?

Keywords Acoustic cues · Auditory memory · Auditory sketching · Perceptual learning · Psychomechanics · Resynthesis · Reverse correlation · Textures · Timbre

T. R. Agus (✉)
School of Arts, English and Languages, Queen's University Belfast, Belfast, UK
e-mail: t.agus@qub.ac.uk

C. Suied
Département de Neurosciences et Sciences Cognitives,
Institut de recherche biomédicale des armées, Brétigny-sur-Orge, France
e-mail: suied.clara@orange.fr

D. Pressnitzer
Laboratoire des Systèmes Perceptifs, Département d'études Cognitives,
Paris Science & Lettres – PSL University, Centre national de la recherche scientifique,
Paris, France
e-mail: daniel.pressnitzer@ens.fr

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_3

3.1 Introduction

In everyday life, there is a constant barrage of auditory events from a wide range of different sources even in the most serene of environments. Reaction to those sounds could vary from disinterest to excitement to fear depending on what the sounds are thought to be. Passing cars, passing trains, human voices, babies' cries, buzzing flies, fighting crows, footsteps, noisy fridges, and tapping on a computer keyboard are just a few sounds that were audible as this sentence was typed. It is widely believed that timbre plays a key role in the recognition of these sounds (even though pitch and loudness would certainly provide useful additional constraints in many cases). The identification of such a vast array of sounds takes place with such apparent ease that this key aspect of auditory perception could easily be taken for granted. This chapter describes experiments that attempted to better characterize and understand the recognition and identification of timbres.

Timbre is multidimensional in the sense that timbre perception is not expected to be described in terms of a single value with all sounds lined up in order from one extreme to the other. Timbre may not even be a single monolithic perceptual attribute. There are many different behavioral uses of timbre (Siedenburg and McAdams 2017). The aspects of timbre appreciated in music (see McAdams, Chap. 8) may not be the same as those upon which timbral differences are judged (see McAdams, Chap. 2), which may in turn be a different timbre from that which signals that a dog has barked. This chapter specifically concerns timbral perception as it relates to the recognition of sounds.

It may seem unnecessarily complex to single out a type of behavioral task to understand a perceptual feature. Even for a classic and respectable perceptual attribute such as pitch, it is quite likely that the task asked of the listener will change the pitch cues being used (McPherson and McDermott 2018). Therefore, adopting an operational definition of timbre, for the sake of this chapter at least, may ultimately prove useful. Here, timbre is defined as that attribute of sound perception that allows recognition of a sound, when pitch, loudness, duration, and spatial position are not informative. This review is also focused on particular everyday sounds, such as musical instruments, environmental sounds, auditory textures, and the human voice. Studies focused on speech have been omitted in order to avoid linguistic, semantic, and social complications (but for a review of talker recognition, see Mathias and Kriegstein, Chap. 7).

Many of the experiments reviewed here involve *identification*, in the sense that listeners are expected to be able to report what the sound is verbally, whether it is a trumpet or a bouncing ping-pong ball. The word *recognition* implies a more general activation of auditory memory, manifesting either through a sense of familiarity or an appropriate behavioral response (McAdams 1993; Yonelinas 2002), such as walking to the front door after hearing a door bell (see also Siedenburg and Müllensiefen, Chap. 4). Pre-lexical recognition, similar to implicit memory, seems more representative of interaction with sounds in everyday life, as sounds rarely need to be named outside of laboratory tasks. However, to the extent that the listeners would know the names for the sounds (e.g., for experienced musicians and familiar instruments), identification and recognition can be treated as equivalent. Nevertheless, identification results may diverge from recognition when an experiment is particularly demanding on the precision with which sounds must be labelled. Distinguishing trumpets, cornets, and trombones is more challenging a task than one in which they could all be labelled “brass” (see also Sect. 3.1).

A holy grail of research in timbre recognition would be to understand what acoustic cues allow listeners to infer the sound source and to determine which acoustic cues are necessary and sufficient for a listener to correctly identify the source. Such a final answer is not yet available. More often, there is evidence pointing toward general categories of cues such as those found in the spectral or temporal domains or in the start or the middle of a sound. Moreover, a pervasive thread is that the cues that are used appear to be affected by learning and the task at hand, suggesting that the links between acoustics and perception may be hard to establish unequivocally, not just because of experimental challenges but also because timbre perception for recognition may be a moving target—a possibility that is entertained in the later parts of this chapter.

Given the wide range of cues that have been investigated for a wide range of sounds, this short tour through the relevant literature is structured according to methodology, highlighting what each method has revealed about timbre recognition so far (Fig. 3.1; Sect. 3.3). Important findings are discussed along with the limitations in the interpretation of these results with the aim of building a picture of the cues that listeners seem to use or seem able to use when recognizing sounds. As many of these techniques have been developed or significantly extended within the last decade, the promise of these methods is as exciting as their early results.

The most direct approaches involve explicitly naming a range of sound sources (*verbal labeling*; Sect. 3.1) or describing properties of the sound-source (*psychomechanics*; Sect. 3.2). To focus the searchlight on more specific features, several

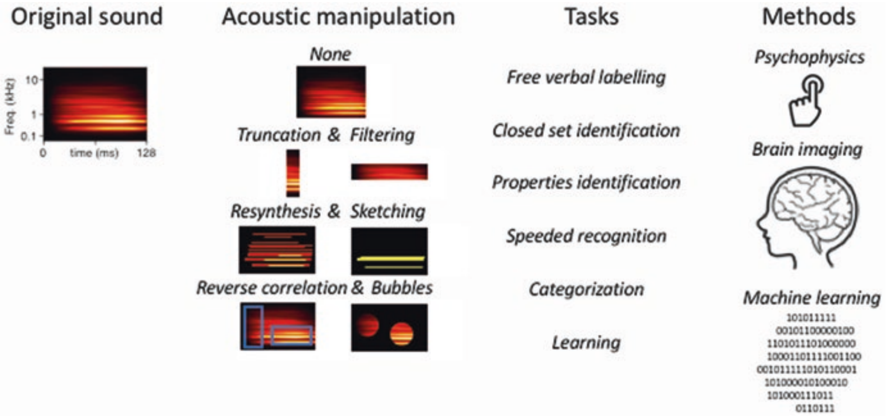

Fig. 3.1 Experimental types reviewed in this paper, all of which can contribute to knowledge of timbre perception

studies have probed the recognition of sounds that have been reduced (*truncation* and *filtering*; Sect. 3.3) or otherwise reconstituted with deliberate information loss (*resynthesis* and *sketching*; Sect. 3.4). Other methods promise to indicate the cues used for recognizing specific sounds or categories of sound with fewer explicit hypotheses (*reverse correlation* and *bubbles*; Sect. 3.5). To circumvent, to some extent, the effect on recognition performance of the unavoidable interindividual differences with auditory experience, other methods observe how previously unheard artificial sounds are recognized to map how recognition evolves over the course of a sound (*time course of recognition*; Sect. 3.4) or over a period of learning (*learning sounds*; Sect. 3.5). Neural imaging provides an alternative viewpoint on recognition (Sect. 3.6.1) and machine learning is discussed as a method of exploring the availability of cues that human listeners could use (Sect. 3.6.2). First, consider the theoretical question that organizes the experimental findings in the subsequent section: Are sounds recognized through broad dimensions, unique characteristics of particular sounds, or a mixture of both (see also Pressnitzer et al. 2015; Siedenburg and McAdams 2017)?

3.2 Continuous Dimensions Versus Unique Features

Timbre as a percept is certainly multidimensional in that sounds cannot be ordered in terms of a single, continuous "timbre line." Alternative representations could obviously include multiple dimensions (planes, open-ended cubes, or higher-dimensional spaces). However, more categorical representations are also to be considered (the presence or absence of a feature, with a very large number of features being considered). A mixture of the two is also possible.

Uncovering a small number of components that can account for as much timbre perception as possible has been a long-term project of timbre research (see McAdams, Chap. 2). This has been explored generally in terms of how similar pairs of sounds are rated to be. Based on these similarity data, there are statistical methods that can position each sound in a *timbre space* (Elliott et al. 2013) in which similar sounds are close together and dissimilar sounds are far apart. The dimensions uncovered in these studies are then sometimes assumed to underlie other aspects of timbre perception, such as identification of sounds (e.g., Ogg et al. 2017). But this may or may not be the case, since different tasks may use different cues even if both tasks are discussed under the umbrella of timbre.

Admittedly, musical instruments that are most likely to be confused are those that are also rated as similar (Grey 1977), and the same is found for environmental sounds (Gygi et al. 2007). Nonetheless, extrapolating this result to dissimilar sounds would make a strong claim that the perceptual features on which dissimilarity is judged may be the same features that are used to recognize individual instruments in isolation. For instance, speech understanding and speaker identification both operate on the same acoustic signal but likely on different cues (Formisano et al. 2008). To further illustrate this through an example in the visual domain, color is a

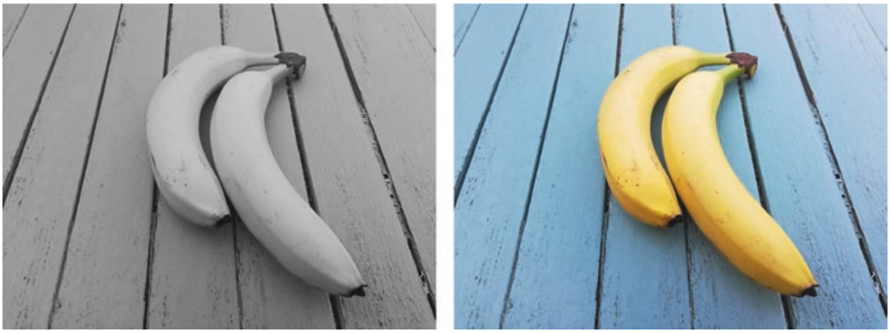

Fig. 3.2 Although bananas are stereotypically yellow, they are easily recognizable without color-based cues. Although color is optional for rapid visual recognition (Delorme et al. 2000), colors still contribute to the similarity or dissimilarity of two images. Analogously, there could be a disjunction between cues that contribute to timbral dissimilarity and those that contribute to recognition

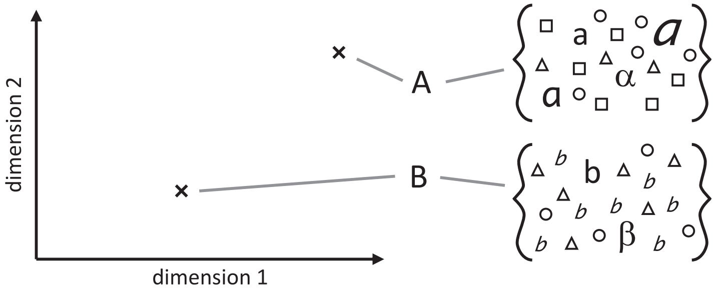

Fig. 3.3 Schematic illustration of two extreme accounts of how two sounds (*A* and *B*) are recognized: (1) All sounds are positioned in a low-dimensional space (*dimension 1* and *dimension 2*) and recognized on the basis of their position in it (*left panel*); (2) in all sounds, a small subset of all the possible features is detected (represented by *symbols*, *letters*), including some esoteric features by which they can be recognized (*right panel*). (Modified from figure 2 of Pressnitzer et al. 2015)

highly salient visual cue, yet a banana remains recognizable in black and white (Fig. 3.2). In fact, when color is removed from photographs, the identifiability of everyday objects is neither impaired nor even slowed (Delorme et al. 2000).

Another possibility would be that timbre identification is mediated by a larger number of cues (Fig. 3.3), some of which are diagnostic of specific sound categories. For example, the specific differences in the envelopes of the harmonics of the sound of a cornet indicate its “brassiness” (Beauchamp 1975). There could be a large number of features, each found only in a small number of sounds or even a single recognizable sound. This is what distinguishes our use of *features* versus *dimensions*. Dimensions are common to all sounds, and there are a small number of them, whereas features may be unique to a class or even a single sound, and there are many of them (Pressnitzer et al. 2015). A mixture of the two ideas is possible. Indeed, inclusion of "specificities" for some instruments has been observed to improve fit to multidimensional models of the similarity data itself (McAdams et al. 1995).

On the neural level, accounts of the response of the auditory cortex typically involve larger features sets, such as the spectrotemporal fields observed in the auditory cortex and generalized for computational models (see Sect. 3.6.2) (Patil et al. 2012). Even these represent a broad set of features whose form is yet limited. These stimulus sets provide a dense representation in the sense that individual sounds may trigger a large proportion of the feature detectors albeit to different degrees. Hromádka and Zador (2009) provokingly argue that cortical activity is more sparse with large numbers of neurons in the auditory cortex remaining quiet for long periods of activity interrupted only occasionally by bursts of activity that are precisely timed to the appropriate stimuli. In theory, such sparse codes bring computational and efficiency benefits (Hromádka and Zador 2009). Whereas it is one role of psychoacousticians to find general patterns in the auditory processes they study, the potential complexity and variety of cues that could be used for auditory recognition should also be kept in mind.

3.3 Experimental Methods for Observing Timbre Recognition

3.3.1 Direct Verbal Labeling

Psychoacoustic methods often boil down to the presentation of a sound followed by a simple question to the participant. For the recognition or identification of a sound, the most obvious question to ask is "what is the sound?" and see how the answers vary with the stimuli. The sets of stimuli selected have ranged from musical instruments to environmental sounds. Although Western musical instruments have little relevance to most people's everyday sound-recognition needs, except presumably for Western musicians, single notes played by instruments have become somewhat of a staple for timbre research. This is likely because they form an easily recordable, closed set of named sounds that can be controlled in terms of their nontimbral auditory properties (e.g., pitch, loudness, and duration).

Berger (1964) asked thirty musicians from a wind band to pair the sounds of ten different wind instruments with their ten names. Some instruments were recognized more reliably than others—29 of the 30 listeners correctly identified the oboe, whereas only 13 correctly identified the flute. This surprisingly poor result for such a well-known instrument could stem partly from the note selected (F4, 349 Hz, at the lower end of treble clef), which is unusually low in the flute's range. A flute sound may be less recognizable at its lower extreme or low notes may be considered less stereotypical of flutes. This specific example serves to illustrate some of the pitfalls of the direct-labeling method.

Berger represented all the listeners' responses as a confusion matrix, counting each of the possible responses for each of the possible instruments. He noted that there were some understandable confusions: The trumpet was confused with the cornet, alto and tenor saxophones were confused with the clarinet, and lower-brass instruments were mutually confused. The challenge of interpreting errors between similar instruments is discussed further in Sect. 3.2. Not all confused instruments were physically similar, however. There was a sizeable number of responses confusing the trumpet with the saxophone in addition to confusions between trumpet and cornet.

The closed response set of Berger seems perhaps not truly representative of day-to-day timbre identification in which there may be less context and certainly not an explicit closed list of possible labels. However, in practice, for a musician hearing an instrumental sound, it is likely that even open responses, not explicitly constrained by a pre-defined list, would nonetheless come from a relatively closed list of familiar instruments.

In other similar studies with broadly similar stimuli, results were highly variable in terms of accuracy, ranging from 33% (Wedin and Goude 1972) to 84% (Grey 1977). In the latter study, listeners benefited from multiple presentations of each instrument throughout the experiment, but even without this practice, Grey's listeners achieved 60% accuracy. There is no doubt that direct verbal labelling is highly valuable as a direct quantitative measurement of timbre recognition, but the absolute values of these studies may be difficult to fully interpret because of the large effects that procedural details seem to have on the results: the choice of the stimulus set, the response set, and listeners' past experiences. Thus, the direct labeling approach has often been complemented by manipulations of the experimental sounds (see Sect. 3.4).

Berger (1964) observed that if confusions within groups of inarguably similar instruments were marked as correct, (e.g., for lower-brass instruments), the 59% accuracy they had observed increased to 89%. In a sense, Berger's reanalysis runs a thought experiment as to what might have happened if only one instrument from each confusable group had been included in the set of stimuli. This illustrates again what a major difference seemingly minor choices about stimuli can have on the overall results. An alternative approach could be to accept more general answers about categories, such as whether the instrument is a brass or string instrument.

Where listeners are asked to categorize according to broad categories, such as musical instruments, human voice, or environmental sounds, they could adopt a range of strategies. At one extreme, they could identify the individual sound as before and then decide to which category it belongs. For example, they could identify a trumpet and then categorize it as a brass instrument. At another extreme, they may be able to categorize the sound correctly without even recognizing the instrument. For example, relatively few people would be able to identify the sounds of the dulcimer, bombarde, and erhu correctly, but most would recognize all three as musical instruments. Thus, when categorizing sounds into broad categories, listeners may use cues that uniquely identify the instruments or cues that are more broadly indicative of the category. Without attempting to summarize the vast literature on categorical perception, the sounds themselves may be more or less prototypical of a category (Rosch and Mervis 1975).

Arguably, all verbal identifications fall somewhere on a spectrum between a unique identifier and a broad category. Even discrete categories of instrument, such as "oboe", can be represented by a variety of sounds generated by different oboes and oboists. In Grey's (1977) study, listeners successfully distinguished two recordings of oboes, albeit with many confusions.

A sound could also be described in terms of its perceptual properties, such as bright, buzzing, and harsh. This could identify the sound in the sense of allowing us to distinguish it verbally from others. However, these semantic labels seem distinct from recognition and identification, as defined in Sect. 3.1, in that they can be used to describe sounds that a listener has not heard before, does not recognize, and could not identify. Semantic labels are discussed in more detail by Saitis and Weinzierl (Chap. 5).

There may be borderline cases where semantic descriptions also identify a sound. For example, if a sound is described as brassy, that perception may well use some of the same cues as would allow us to identify a sound as that of a brass instrument. In particular, if a listener is able to describe a sound as "glass being struck by wood but not breaking", then this amounts to identifying the sound without having a specific name for it.

3.3.2 Psychomechanics

The study of a listener's ability to describe the physical properties of the sound source is a field in itself. *Psychomechanics* studies the relationship between the physical properties of an object (its size, material, and shape) and the perception of its physical properties based on sounds that it makes.

At first glance, psychomechanics seems to add a step beyond psychoacoustics, which only aims to find relations between the emanated sounds and perceptions of them. However, Lakatos et al. (1997) note that a single sound source could emanate from a wide variety of sounds, so in order to recognize a sound source, it would be efficient to focus on invariant cues, cutting through the complexity of the acoustic medium. This is one motivation for studying psychomechanics. The wide range of psychomechanical results are reviewed by McAdams (Chap. 2), all of which are relevant for sound source recognition and identification insofar as it is sufficient to identify the sound's properties.

The suitability of the psychomechanical approach may depend on the category of sound under question. All but the simplest acoustic couplings can be highly nonlinear and hard to predict (Lakatos et al. 1997). For instance, although most people are proficient users of the human voice, few have any particular insight about the acoustic mechanisms involved in vocal production. Nevertheless, in some situations, such as an object being dropped, it may be more useful to identify summary information about the physics of the incident (size, material, broken, or bouncing) than to be able to categorize and label the sound. The relative ease with which these impact sounds can be modeled and manipulated may make them give up their perceptual secrets more readily, and they have proven to be a fruitful sound category for psychomechanics (for a review, see Giordano and McAdams 2010).

3.3.3 Truncations and Filtering

One way of pinpointing the acoustic information that is useful for recognition is to remove parts of the sound and see if it can still be identified. This can be implemented by truncating parts of the sound in the time domain—omitting the start, middle, or end of the sounds—or by the equivalent procedure in the frequency domain, which is filtering.

The information conveyed at different stages of a sound have received considerable discussion. In most musical instruments, an unstable start (the onset) builds up into a relatively stable pattern (the steady state) that can be sustained or die away. For bowed string instruments, the irregular scratch of bow on string is the onset that quickly gives way to a more regular periodic pattern in the steady state. For the timpani, some of the modes of vibration in the onset die away within tens of milliseconds, leaving more durable modes of vibration to dominate the steady state (Benade 1990). The analysis of a periodic sound's partials through a Fourier analysis initially emphasizes the steady-state center of periodic sounds (e.g., the "tones" of Helmholtz 1877; for a review see Risset and Wessel 1999); whereas the ever-changing onsets of realistic sounds defy such clean analysis (see Backhaus 1932 for a valiant effort).

Although the onset of sound is often credited with particular importance in the recognizability of sound (e.g., Iverson and Krumhansl 1993), the pattern of results is much more complex. Generally, the accuracy of identification declines somewhat when onsets are removed (Saldanha and Corso 1964; Suied et al. 2014) but not catastrophically. Some instruments are affected more than others. Saldanha and Corso's (1964) double bass and English horn were confused if their attack transients were removed, and the middle portion of Berger's (1964) alto saxophone was often identified as a trumpet. For Elliott (1975), string instruments were mostly identified as the oboe without their onset, but most instruments were more likely to be identified correctly than as a different specific instrument. Wedin and Goude (1972) generalized that the instruments that survived removal of the onset were those that were recorded with vibrato. Onsets presented in isolation seem to be relatively identifiable, although not as much as the intact sounds (Clark et al. 1963). These combined results point in the direction of fluctuations as being important for identification, less so the steady state (where a steady state is achieved). This could be because the steady state provides rather limited and unreliable information about specific frequencies, whereas sharp transients or fluctuations convey information about wider ranges of the spectrum.

Another question tackled through truncations is how long a sound must be to be recognized. This provides an upper bound for the duration of the acoustic cues on which the recognition is based. Seminal studies using this technique were reported by Gray (1942) for speech vowel sounds and by Robinson and Patterson (1995) for musical-instrument sounds. Surprisingly, results showed identification above chance for durations that were even shorter than that of a single repetition of the periodic waveform. Bigand et al. (2011) asked listeners to categorize a more diverse set of sounds, such as spoken voices, instrumental music, or environmental sounds (including water bubbling, a cigarette lighter, and a foghorn). The voices and instruments were categorized above chance within 20–30 ms, whereas the environmental sounds required 50 ms. These durations are similar to the categorizations (at 25 ms) observed by Ogg et al. (2017) using the same categories with a different set of stimuli. Suied et al. (2014) used a more focused sound set, all periodic, in which pitch cues were counterbalanced across the set and could not be used for recognition. With fewer differences between the sounds in the categories, above-chance categorization of sung vowels or musical instruments was observed, counterintuitively, at much shorter durations, as little as 2–4 ms. Such short truncations only leave a snapshot of the spectrum available and a rather blurred indication of it at that (see left panel of Fig. 3.4) (Occelli et al. 2016). The best recognized categories also seem to depend on properties of the stimulus sets, either the voice (Bigand et al. 2011; Suied et al. 2014) or the more spectrally stable instrument (Ogg et al. 2017). Thus, different cues may be available at different durations, and whether these cues allow correct categorization could depend on the context in which the sounds are

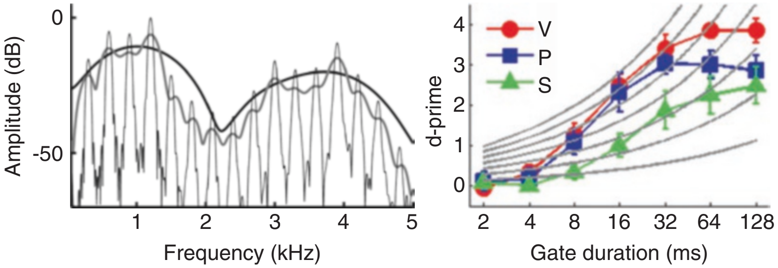

Fig. 3.4 Distinguishing sounds. **Left panel:** Fourier transforms of a recording of the vowel sound /a/, sung at pitch D4 (294 Hz), truncated with raised-cosine windows. For a 32 ms sound snippet (*thin black line*), the harmonic structure of the sound is visible in the form of regular peaks and their heights outline a detailed spectral shape. As the snippet gets shorter (8 ms, *thick gray line*; 2 ms, *thick black line*), frequency resolution is lost and only broader spectral features are available. In spite of the vast differences among all of these sounds, listeners made "above chance" decisions when asked to categorize them. **Right panel:** The ability of listeners to distinguish short snippets of voice (*red*), percussion (*blue*), and string instruments (*green*) from seven other musical instruments is plotted by gate duration (ms) and d-prime (a statistic related to the perceptual distinctiveness of the two conditions). The *gray lines* indicate the rates of improvement that would be expected if listeners were only taking advantage of repeated opportunities to hear cues that were available at shorter durations. The steeper slopes in the behavioral data indicate that additional cues have become available. (Both panels modified from Suied et al. 2014; used with permission from AIP Publishing)

presented. In summary, listeners are able to make use of rough spectral cues that are available within milliseconds, but to distinguish between a wider set of sounds more robustly, a wider range of cues available by 25–50 ms may be required. The build-up of cues over time is discussed further in Sect. 3.4.

In the spectral domain, Gygi et al. (2004) explored the identification of filtered versions of seventy everyday environmental sounds obtained from a sound effects library, including everything from airplanes, baby cries, and the opening of a beer can, to waves, windshield wipers, and zippers. Listeners were trained with feedback to the point of 95% accuracy at identifying the sounds before being asked to identify the sounds without feedback. Some sounds were more affected by filtering than others. Perhaps predictably, thunder became difficult to recognize when high-pass filtered with a cut-off of 300 Hz. Sirens and laughter were highly recognizable however destructively they were filtered. Nonetheless, the authors were interested in the frequency regions that were globally the most useful. Approximately 80% accuracy was obtained across all sounds and listeners when sounds were band-passed to a 1200–2400 Hz frequency region. However, the sounds were also highly identifiable when high-passed well above this region (over 70% accuracy with a cut-off at 8000 Hz) or low-passed well below (over 50% identification with a cut-off of 300 Hz). This shows that the identifiability of a wide range of environmental noises is surprisingly robust to spectral distortions. It also shows that cues for identifiability should be found across a wide frequency range.

Truncations and filtering are inherently destructive forms of manipulating the stimuli. Only a decrease in identifiability would be expected with either technique; the relative importance of different times or frequency regions can only be inferred from the extent to which their removal affects performance. Also, removing an acoustic cue is not simply omitting it but rather replacing it with an alternative cue. For instance, cutting the onset amounts to introducing a sharp onset. Such alternative cues may themselves affect a listener’s responses. To manipulate the presence or absence of cues more generally, resynthesis has been used as a complementary tool.

3.3.4 Resynthesis and Sketching

Resynthesis refers to the process of synthesizing a new sound based on the measured acoustic features of a pre-existing sound. If one can control the presence of a cue in a sound with reasonable independence from other cues, then testing the role of that cue for auditory identification becomes considerably easier. *Sketching*, a notion introduced more recently, refers to a subcategory of resynthesis tools that aim to use a relatively small number of features.

Cues that have been manipulated include the spectral centroid (Wun et al. 2014), the timing of repeated sounds (Warren and Verbrugge 1984), and the envelopes of individual harmonics (Grey and Moorer 1977). Many studies use resynthesis to test the extent to which a sound is similar to the original (e.g., Grey and Moorer 1977; McAdams et al. 1999).

Vocoders are a form of resynthesis of particular interest because they preserve and destroy cues in ways that are analogous to cochlear implants (Shannon et al. 1995; see Marozeau and Lamping, Chap. 9). Noise vocoders generally preserve the rough shape of the spectrum of a sound and reproduce changes in the spectrum over time, so many of the broad spectrotemporal cues remain. However, vocoders resynthesize a sound from bandpass-filtered noise by simply changing the level of each noise band. Thus, details of the spectrum that are finer than the vocoder's bands are lost and replaced by those of noise. Any harmonics are replaced by bands of noise, so nearly all pitchiness is lost. A similar principle applies to sine wave vocoders, simply replacing noise bands by fixed-frequency pure tone carriers.

Gygi et al. (2004) also tested whether a range of environmental sounds could be recognized from versions preserved by a noise vocoder. As expected, sounds with strong pitched components were among those least well identified, including the flute, car horns, and an electric saw. More surprising is that some of the sounds could be identified correctly after processing with a six-channel noise vocoder. Specifically, 36% of sounds could be identified by naïve listeners who had never heard the original sounds. The most readily identifiable sounds included a baby cry, bird song, a clock, and a bouncing ping-pong ball (each 98% accuracy), some of which also have considerable harmonic content, but all of which have distinctive rhythmic patterns. This emphasizes that not all sounds are identified on the basis of the same acoustic cues. Different sounds afford different cues through which listeners might recognize the sound source.

A powerful resynthesis technique has been developed by McDermott et al. (2011), specifically geared toward *auditory textures*. Textures refer to relatively steady sounds of indeterminant length, such as water running or fire crackling. McDermott et al. (2011) extracted a large number of parameters (over a thousand) inspired from a model of human auditory processing with parameters relating to the intensity within frequency bands, the distribution of intensities over time, the rate of modulations within each frequency band, and correlations between those (Fig. 3.5). A noise stimulus was then progressively altered in an optimization process until it reached similar statistics to a target texture sound. The reconstructed sounds were recognized with a high level of accuracy (89%, compared to 96% for the original sounds). Each parameter seemed to play a role in improving identifiability of the sounds; notably, the addition of correlations across frequency bands in the resynthesis model vastly improved recognition from 62% to 84%, relative to within-frequency-band statistics alone. This shows that cross-frequency correlations, in conjunction with more classic auditory statistics, play a role in preserving the acoustic cues that are important for recognition. The precise nature of these cues is unknown, but they point toward a recombination of information from across frequency bands that is otherwise little mentioned and little observed physiologically.

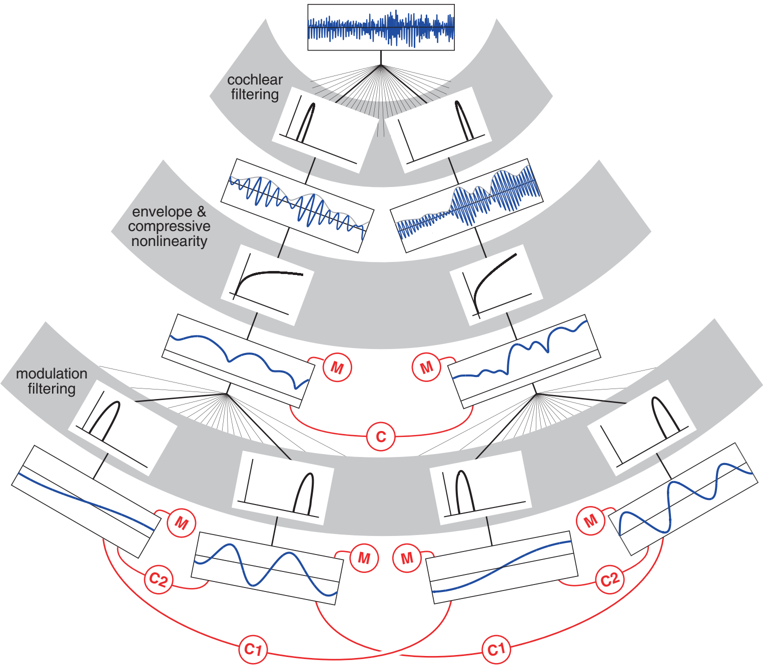

Fig. 3.5 Schematic of the analysis model of McDermott and Simoncelli's (2011) resynthesis method. Sounds pass through a bank of bandpass filters after which envelopes are extracted and compressed. These signals are in turn passed through the bandpass filters of the modulation filterbank. Statistics are analyzed at various stages, including marginal statistics (represented by *M*, which can include the mean, standard deviation, skew, and kurtosis that are not shown) and correlations between various channels (*C*, *C1*, and *C2*). (From McDermott and Simoncelli 2011; used with permission from Elsevier)

An opposite extreme of resynthesis is to test how *few* parameters suffice to preserve recognition. This has been termed *auditory sketching* (Suied et al. 2013b) as a parallel to visual sketches. Visual sketches can be easily recognized despite extreme simplifications of the original image. The potency of visual sketches is thought to be related to their use of cues adapted to the human visual system, such as lines (Cavanagh 2005). The discovery of the equivalent basic cues for audition, if they exist, would be an important step in understanding timbre recognition.

Isnard et al. (2016) generated audio sketches of instruments, birds, vehicles, and voices based on the times and frequencies of peaks of energy,[1] and asked listeners to categorize them in order to explore the importance of the representation of sound from which the sketches were generated. There were two types of sketch. One is acoustic, based on Fourier analysis, and the other is an auditory sketch based on a

[1] Examples of acoustic and auditory sketches are available from https://hal.archives-ouvertes.fr/hal-01250175 (Isnard et al. 2016)

model of the human cochlea (Chi et al. 2005). Only relatively stable sound textures (vehicle engine noises) were better categorized from the acoustic sketches than from the auditory sketches. In fact, the rate at which listeners categorized acoustic sketches of the voice correctly was below chance, suggesting that listeners were actively miscategorizing them. Both bird song and voices were better categorized based on the auditory sketches than acoustic ones. Even with as few as ten features selected, listeners could categorize the sounds of vehicles, birds, and musical instruments above chance, averaged across both types of sketch. More broadly, the technique of sketching could be used in the future to test which combinations of basic acoustic cues are sufficient to allow recognition, or even which acoustic parameters allow the most efficient representations of sounds for recognition tasks.

3.3.5 Reverse Correlation and Auditory Bubbles

When synthesizing or manipulating sounds, there is always a choice as to which features are worth changing and which are worth preserving. Ultimately, the experimentalist decides which distinctions are worth making and testing. The techniques of reverse correlation and its offspring, auditory bubbles, promise to avoid these hypothesis-driven methods and allow listeners' responses to speak for themselves. In reverse correlation, a listener's responses to a wide range of stimuli are correlated with parameters of those stimuli. Such analyses have been used by Ogg et al. (2017), who observed that listeners' categorizations as speech, instrumental music, or environmental sounds were correlated with the measured spectral centroid, spectral flatness, and noisiness of the snippets presented. A risk with post hoc analyses is that the responses may correlate with several factors that covary in the stimulus set. Indeed, it is possible that some correlated features might have had no causative role in the participants' responses. Whereas this is often a risk worth taking for an indication of the cues that listeners relied on, stimulus sets for reverse-correlation experiments are designed to avoid confounding correlations.

The stimulus set designed by Brimijoin et al. (2013) consisted of 0.5 s noises whose spectra varied randomly from trial to trial. Listeners were asked to respond as soon as they heard a specified vowel sound. On average, the stimuli that triggered listeners' responses tended to have more power in some parts of the spectrum than others, resembling the spectral shape of each of the vowels in whispered speech. Moreover, the stimuli immediately before the one that triggered the response tended to have the opposite pattern with less power where the vowel would normally have peaks. This shows that listeners were not just responding to the absolute spectrum of the latest noise stimulus but also to its contrast with the preceding noise. This highlights yet another little-mentioned possibility, that there is sensitivity not just to the spectrum of a sound but to its changes compared to an acoustic context (cf. Sidwell and Summerfield 1985).

The method of auditory bubbles (Venezia et al. 2016) combines the techniques of auditory sketching with reverse correlation. The original bubbles method, as used in visual studies, consisted of circular regions of an image shown while the rest of the image was blacked out (Gosselin and Schyns 2001). A typical finding has been that the parts of the facial images that best correlated with participants' responses depended on whether they were trying to judge the gender or the expressiveness of the faces presented. The technique was developed as a more efficient alternative to reverse correlation, which typically requires several thousand trials to provide meaningful results.

In terms of timbre recognition, a bubble methodology has been implemented by Thoret et al. (2016) for instrumental arpeggios and applied to individual sustained instrumental sounds (Thoret et al. 2017). Listeners were asked to identify instruments based on resynthesized versions generated from a relatively narrow range of their spectral and temporal modulations, effectively blurring and sharpening the spectrogram to emphasize certain rates of change in both the spectral and temporal directions. Based on the pattern of identifications by listeners, some regions contributed to a larger proportion of correct responses than others.

Focusing on sustained instruments (Thoret et al. 2017), identification overall was most reliable based on lower rates (less than 30 Hz in the temporal domain and over 18 cycles/Hz in the spectral domain), but the most useful regions varied by instrument. For example, the cello was best recognized on the basis of its slower modulations (less than 18 Hz) and broader spectral details (less than 20 cycles/Hz) while the saxophone was recognized on the basis of faster modulations (10–30 Hz) and its broader spectral features (less than 15 cycles/Hz). A similar analysis focused on situations where instruments were misidentified. The patterns of confusions were consistent with preserving parts of the modulation spectrum that were normally used to recognize the wrong instrument of the pair: emphasizing the faster modulation rates of the cello (18–30 Hz) made it more likely to be misidentified as a saxophone and emphasizing lower modulation rates of the saxophone (less than 12 Hz) led to misidentifications as a cello. The bubbles method, as implemented in those studies, suggests that identification and misidentification is mediated by spectral and temporal shapes at specific scales that can be described in terms of spectral and temporal modulations. Only a single bubble was used, so extending the technique to multiple bubbles may lead to further insights about how information from different regions is combined.

Reverse correlation and its bubbles-based variant provide tools through which any feature space, however large, can be explored, in theory. Therefore, they add a powerful weapon to the psychophysicist's toolbox, which is otherwise better kitted out to compare a small number of carefully controlled conditions.

3.3.6 Interim Summary

Returning to the quest of finding the cues that mediate recognition of everyday sounds, the methods discussed so far have not led to the Holy Grail but they have motivated a journey. The truncation and filtering studies (Sect. 3.3) show that the onset has particular importance for some musical instruments but not all. Likewise, in the spectral domain the key cues for environmental sounds are found in different frequency ranges for different environmental sounds. Thus, it will be difficult to uncover the cues of recognition with a one-size-fits-all strategy, and it may be necessary to study what gives individual sounds their properties.

A key lesson from the resynthesis and sketching methods (Sect. 3.4) is that listeners can sometimes recognize sounds that are highly distorted, preserving only the rough shape of a sound (in vocoding) or major landmarks in a sound (in sketching). This could come about from the versatility of listeners, suggesting that familiar sounds can be recognized from a wide variety of cues as necessary. While making timbre recognition a more challenging topic for the experimenter, this trait could be particularly useful in less ideal listening conditions when masking or reverberation might render some cues useless while leaving others available.

Reverse correlation and auditory bubbles (Sect. 3.5) pointed toward features that can be clearly seen in a spectrogram as power that varies in time and frequency. While this supports a traditional view of auditory perception in terms of its emphasis on the power spectrum, this could simply reflect the feature spaces that were explored through these methods. If listeners are able to use a wide range of cues opportunistically, then it should be expected that listeners use the types of cues that are made available to them within a given experimental setup.

3.4 Observing the Time Course of Recognition

Recognition does not happen in an instant. The information available in a sound builds up over time and it takes the listener additional time to process that sound and respond. Observing the build-up and processing of this information gives some insight into the mental processes and acoustic cues involved in recognition.

For example, in the truncation experiment of Suied et al. (2014) that measured the categorization of the human singing voice, string instruments, and tuned percussion (see Sect. 3.2.3), it was noted that although categorization above chance levels was possible at 4–8 ms, performance continued to improve as the duration of the stimulus increased (see right panel of Fig. 3.4). An increase in performance could be due to hearing additional features but would also be expected merely from the additional opportunities of hearing the same short cues multiple times during the stimulus. This could be envisioned as making regular guesses each time a cue was or was not heard with the final answer representing the majority opinion. The majority opinion would be expected to be more reliable than each individual guess. This intuition has been formalized mathematically using signal-detection-theory models (Viemeister and Wakefield 1991), predicting that sensitivity would double each time the signal duration quadrupled (or more generally, that d' would increase in proportion to $\sqrt{T}$, where T represents the duration of the stimulus and d' is a common measure of sensitivity). The gray lines on the right panel of Fig. 3.4 show the trajectories of improvement over duration predicted by this model. In practice, listeners' performances improved with duration that exceeded this amount over the range 4–32 ms, depending on the target category. This indicates that additional cues for recognition become available across this time scale, which is consistent with the results from categorization tasks using more diverse stimuli (Bigand et al. 2011; Ogg et al. 2017). These could take the form of more detailed representations of the spectra or cues that involve changes over time.

In situations where such short stimuli can be recognized, there is the opportunity to see the rate at which listeners can categorize sounds. Suied et al. (2013a) asked listeners to pick out snippets of human voice (16 ms or 32 ms long) from a rapid sequence of equally short snippets of musical instruments in the same octave range of pitches. Performance generally decreased with increasing presentation rates, but listeners performed above chance up to 30 sounds per second (at which rates only the 16 ms stimuli could be presented). Effects of streaming or nonsimultaneous masking might come into play at such rates. However, performance did not deteriorate when pitch was fixed across snippets, as would be expected if either streaming or nonsimultaneous masking was the limiting factor. Rather, a limit of timbre perception seems to be reached when processing a different sound snippet every 30 ms.

Whereas that duration puts an upper limit on the distinct sounds that can be processed within a second, it almost certainly takes longer to recognize an individual sound from the time the sound is presented to the point that it is recognized. An upper limit can be put on this amount behaviorally by asking listeners to react as quickly as possible when they recognize a sound, bearing in mind that this includes the time taken to respond. When asked to simply let go of a button as quickly as possible following any sound, listeners took roughly 300 ms to respond (Agus et al. 2012). Asking them to respond only to the human voice while ignoring a range of instruments added another 150 ms. This was faster than an equivalent task with string instrument targets (105 ms slower) or even tuned percussion instruments (55 ms slower than the voice). The question of whether the human voice is processed advantageously over other timbres irrespective of the stimulus set is a complex one (see Siedenburg and Müllensiefen, Chap. 4), but when such a behavioral advantage is seen, there is an opportunity to investigate the acoustic basis of the advantage.

Agus et al. (2012) also presented resynthesized *auditory chimeras* that combined aspects of voice, string, and percussion instruments. These were generated by imposing the long-term average spectrum of one sound on another sound in such a way as to preserve details in its temporal spectrum, including temporal fine structure and relative spectral fluctuations (see Fig. 3.6; audio demos available at http://audition.ens.fr/chimeras). If the voice's spectrum, for instance, was the key to the faster responses, then listeners should be able to respond as fast to the chimeras that preserved the voice spectra as for the original voice itself. This is not what was observed. In fact, selective responses to the auditory chimeras were roughly as slow as those for the original instrumental targets whether or not the chimeras preserved either the temporal structure or the spectrum of the voice. This indicates that the faster responses do not stem from either the spectrum or the temporal structure of

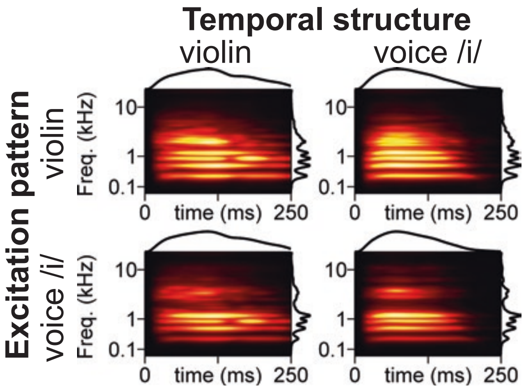

Fig. 3.6 Some of the natural stimuli (*top-left* and *bottom-right* panels) and auditory chimeras (*top-right* and *bottom-left* panels, voice /i/) used by Agus et al. (2012). Each panel shows how power varies over time and frequency (*colored plots*) as well as the envelope (*top of panel*) and the long-term average spectrum (*right of panel*). Each chimera preserves the long-term average spectrum of one of the natural stimuli and the temporal structure of the other. (From Agus et al. 2017; reprinted under the Creative Commons CC-BY license)

the voice alone but from joint spectrotemporal cues, such as a time-varying spectrum or a combination of separable spectral and temporal cues.

In combination, these studies suggest that simple cues, such as the spectrum, are used when the task demands it, but additional useful cues build up over time when they are available. In a naturalistic sound, the combination of cues that trigger recognition may not fit neatly into either the spectral or the temporal category.

3.5 Learning Sounds

Up to this point, this chapter has treated the recognition of sounds as a relatively stable phenomenon, meaning that it is possible to probe established relations between the sound and the labels associated with it. Presumably, these associations were originally learned, perhaps early in development. If one can observe the learning of new sounds in the laboratory, auditory recognition may be probed from new angles because: (1) the types of features that are most easily learned may indicate the types of features that have been learned for more familiar sounds; and (2) the stimuli that listeners learn can be designed to have better acoustic controls than occur in natural stimuli.

The rapid learning of a complex novel stimulus was observed by Agus et al. (2010), specifically, learning of an arbitrary snippet of white noise. Listeners demonstrated their learning of the noise indirectly through improved performance in a difficult task. The task involved white noises that were either reference noises that they heard regularly within a block of trials or fresh noises that they had never heard before. Specifically, they were asked to detect when a one-second noise was formed from two identical half-second noises. Rapid and robust learning of the noises occurred within a few presentations, leading to near-perfect repetition detection for roughly a third of the reference noises (with no learning observed in the remaining two-thirds). The learning seemed to be idiosyncratic in that a noise that was easily learned by one listener was not particularly easy to learn for other listeners. Follow-up tests showed that learning transferred to frequency-shifted stimuli up to a third of an octave (but not as much as half an octave) and also transferred to reversed stimuli.

These results suggest that the learned features were relatively local in frequency and did not extend over long periods of time. A variant of the experiment also confirmed that learned features for noise were confined to relatively short periods of time (Andrillon et al. 2015). Finally, learning with rapid and robust characteristics similar to that of white noise was observed with a similar experimental procedure that used random tone clouds as stimuli (Kumar et al. 2014). However, Kang et al. (2017) showed that listeners could also learn sounds deliberately stripped of the rich spectrotemporal variations of noise, but this learning did not extend to reversed click trains, highlighting that longer temporal patterns were learned. So, it seems that several types of cues—spectral, spectrotemporal, or purely temporal—can be efficiently learned by the auditory system.

In a different paradigm that used textures instead of white noise, McDermott et al. (2013) investigated the listener's ability to distinguish different snippets of the same or different auditory textures. Short snippets were easier to distinguish from one another than longer ones when the snippets were all drawn from the same texture. The opposite was observed when distinguishing snippets from different textures. This was interpreted as an indication that, beyond a certain point, listeners retained only summary statistical information about the cues that were present in the sound, and such summary statistics were the cues used to recognize textures. For different sound sources, the statistical differences would build up over time and lead to improved performance, whereas for snippets of the same auditory texture, the summary statistics would increasingly converge and lead to poorer performance.

These two sets of results seem to point toward two distinct learning mechanisms for recognizing a sound source: local features or summary statistics. The boundary between the two mechanisms is blurred when the features are extended over time or when the statistical interpretation is applied over a sufficiently short period of time.

Overall, these results suggest impressively fast and robust learning mechanisms for timbre cues that can adapt to the characteristics of the sounds to be learned. Where a sound is spectrally rich, such as a noise, listeners may learn local spectral cues. When sounds contain only temporal information, listeners may learn more temporally extended cues. For the many sounds with more complex acoustic structures, such as the human voice or auditory textures, listeners may learn spectrotemporal cues on several time scales. Moreover, the cues may well be idiosyncratic to each listener, depending on past experience. Altogether, the experiments show that the human auditory system is equipped to learn the wide range of sound sources it will encounter throughout a lifetime.

3.6 Neuroscientific Approaches

3.6.1 Neural Correlates of Sound Recognition

The neural representation of timbre in general is the topic of Alluri and Kadiri (Chap. 6), so this section focuses on the studies that specifically extend understanding of timbre recognition and the cues and processes underlying it.

Acoustic confounds between stimuli in different conditions are treated carefully in brain-imaging studies because if low-level acoustic differences broadly affect the pattern of excitation at the cochlea, these differences would likely be reflected in all subsequent parts of the auditory system. Thus, irrelevant low-level cues could masquerade as features that distinguish between categories. Different studies have addressed the risk of confounds with different strategies: making comparisons with diverse conditions that are unlikely to suffer from the same acoustic confounds (Belin et al. 2000), factoring out the acoustic differences that are present (Leaver and Rauschecker 2010), or selecting and processing stimuli to minimize the acoustic differences (Lewis et al. 2004; Agus et al. 2017). With these controls comes the dual risk that a low-level confound is overlooked or that important cues for timbre categorization are deliberately omitted from the design. Rather than preselecting categories, Norman-Haignere et al. (2015) presented a wide range of natural sounds and inferred distinct components that seemed to be underlying their diverse responses.

Broadly, these different techniques report differences driven by various acoustic features at the level of the auditory cortex, including frequency content, pitch, and temporal modulations; whereas higher-level categorizations of the sounds, such as human vocal sounds, may emerge in the nearby regions of the superior temporal gyrus and the superior temporal sulcus along with selectivity to the patterns in speech and music that are built on variations in pitch, loudness, and timbre. Notably, patients who develop difficulty in recognizing sounds are often found to have lesions around these temporal regions as well as in the hippocampus (Tomasino et al. 2015), which is often associated with memory, suggesting that sound source selectivity may result from extensive exposure to familiar sounds.

A powerful method of obtaining acoustically controlled conditions for brain imaging is to use the same sounds for different conditions, changing only the task. Hjortkjær et al. (2018) presented listeners with the sounds of wood, glass, or metal being either struck, rattled, or dropped. In one session, listeners were asked to report the material. In another session, listeners were asked to report the action. As might be expected, sounds that were acoustically similar led to neural responses that were also similar, at least in Heschyl's gyrus and the nearby planum temporale. Specifically, similarity in spectra correlated with similarity of responses in Heschyl's gyrus, which contains the primary auditory cortex. Of particular interest here, this correlation was stronger when listeners performed the material-identification task. The implications are two-fold in that the task affects a listener's representation of the sound (cf. Fritz et al. 2003), and the patterns of neural activity that are correlated with the sound source's material are also correlated with the spectrum, in line with a previous study (Hjortkjær and McAdams 2016). Similar patterns were observed for temporal modulations in both Heschyl's gyrus and the planum temporale with stronger correlations observed during the action-identification task. Thus, the neural representation of sound can be affected by the task the listeners are performing, yet this apparent complication allows insights as to the acoustic cues that underlie these different representations.

Brain imaging studies draw attention to the possibility that different types of sound may excite different regions of the brain to different extents. This opens up the possibility that specialized cues could come into play for important categories of sound. It also becomes apparent that different cues may be emphasized to different degrees. The cue variability with task initially seems to undermine the quest to find which cues subserve recognition, but there remain questions about how flexible this task-related variability is. Does it involve different sets of cues or is it merely a matter of emphasis? Which cues can the task accentuate or de-emphasize? Similarly, if cues come into play for specific categories of stimuli, this pushes the question back a step to ask which cues trigger the involvement of the specialist processing.

3.6.2 Machine-Learning Approaches

Machine-learning approaches can be used to explore what information is available in a given representation of sound or to explore which representations provide enough information to perform a given identification or categorization task. Some computational descriptions of the acoustics relate more directly to human auditory perception than others (see Siedenburg et al. 2016 for a review), so the focus here is on biologically inspired machine-learning studies (but also see Caetano, Saitis, and Siedenburg, Chap. 11).

Coath and Denham (2005) generated sets of feature detectors that responded to different patterns of onsets and offsets within each frequency channel of a cochlear model and trained an artificial network to categorize a small number of different sounds. They showed that a network trained on a small number of speech fragments (short sections of a single talker saying eleven words) was able to generalize, albeit with a limited degree of accuracy, to classify the same words spoken by a large number of talkers with different amounts of stationary noise added and even time-compressed versions of the stimuli. Whereas this characterized the feasibility of classifying sounds already experienced in some form, they emphasized the importance of developing feature sets that could also categorize novel sounds with which the system was not trained, a skill that is surely necessary for everyday listening. They also tested whether training with eleven environmental sounds could be generalized by the network to distinguish speech sounds. Coath and Denham (2005) found that the ability to distinguish the speech sounds remained but in a slightly reduced state. This suggests that a neural network (artificial or otherwise) could develop its own feature space suited to the sound environment to which it is exposed, but the feature space would also capture some useful information about novel sounds.

Patil et al. (2012) trained a "classifier" to distinguish eleven musical instruments based on training with over a thousand real cortical responses recorded from ferrets. The cortical responses were characterized by their spectrotemporal receptive fields (STRFs), which captured typical responses that reflected the frequency content and the rapidity and direction of the change of the frequency content that had triggered the neural activity (for a detailed introduction to STRFs, see Elhilali, Chap. 12). Based on this representation, the eleven instruments were categorized with 87% accuracy, using a large number of recordings of these instruments playing different tones in different manners. This study shows that cortical responses, as characterized by their SRTFs, include features that allow reasonably reliable classification of these musical sounds.

The recorded SRTFs used by Patil et al. (2012) may not have been representative of a typical population of cortical neurons, as they included experimental noise and potential selection biases, so idealized versions of the SRTFs (Chi et al. 2005) were also tested. The machine-learning classifications increased to 99% accuracy with the idealized versions, showing that the richer representations preserved diagnostic features better. Such performance was not observed for simpler machine-learning models that were based on the long-term average spectra (79% accuracy). Patil et al. (2012) also applied the same architecture to a different task: reproducing judgements of timbre similarity collected with human listeners. They found that the same feature set, but a different metric, was able to reproduce similarity judgements. Thus, an account of robust recognition probably involves relatively complex feature sets, for machines or humans alike, and the way the feature set is used could depend on the task.

Newton and Smith (2012) explored categorization of instruments based on their onsets alone using a model cochlea and model neural responses. Their automatic classifier could distinguish five musical-instrument sounds (brass; reed instruments; and bowed, plucked, or struck strings) with a similar degree of accuracy as more established methods that were based on spectral analyses of the whole tone (at least when the classifier was tested using the same set of sounds it had been trained on). When the classifier was applied to a new set of recordings, without additional training, the performance of the whole-tone spectral classifier deteriorated while the onset-based model maintained its level of performance. These comparisons show that there is robust information available in the onsets of instrumental sounds that could be used in combination with the later steady-state portions of the sound.

Deep-learning networks can now be used to explore auditory object recognition without making explicit assumptions about the underlying cues (Luo et al. 2017). If combined with a model of the auditory periphery, these provide some promise of uncovering the optimal cues available for various situations, especially if responses of subsequent stages of processing are well predicted by suitably trained artificial networks (see Kell et al. 2018 for tasks more complex than timbre recognition).

3.7 Summary

The study of timbre recognition appears to be undergoing a renaissance in terms of the flourishing of new techniques being applied to it. As the range of methods has increased, so have the types of cues reported to play a role in recognition. Spectral cues have been extended to contrasts in spectrum (Brimijoin et al. 2013), spectral fluctuations have been subdivided into rich sets of spectrotemporal modulations and their linear interactions (Chi et al. 2005), and the spectral evolution at the onsets of sounds has been extended into ongoing correlations across frequency bands and their modulations (McDermott and Simoncelli 2011). The principle of parsimony in science should bias us toward finding an account of timbre recognition in terms of the smallest feature set possible, but there is the distinct possibility that cues underlying recognition may be irreducibly complex (Agus et al. 2012). In addition, the cues may vary according to task (Hjortkjær et al. 2018) and sound category as defined by learning and experience (Belin et al. 2000). Such complexities perhaps encourage aiming for more focused challenges, such as understanding the cues used to recognize more constrained types of sound sources.

Many of the experiments discussed in this chapter describe efforts to understand cues that are generally useful (e.g., Gygi et al. 2004), and there are fewer experiments that focus on tightly restricted sound categories (e.g., impact sounds) (Giordano and McAdams 2006). The broad brushstroke approach provides general information as to the most important frequency range to preserve when bandwidth is limited, but when designing auditory enhancements (e.g., for hearing aids) it could eventually be more useful to appreciate the microscopic details of the features that ought to be preserved for effortless recognition of individual sounds. Although this line of research seems more targeted, and therefore limited, it seems just as important for a clearer picture of how each of many individual cues is important for timbre recognition.

A limitation of most of the current timbre recognition literature is that it primarily focuses on cleanly recorded sounds presented in isolation (for an exception, see Gygi and Shafiro 2011), unlike the masking-rich environments that are more typical of everyday life. Different background sounds could render different cues useless, perhaps providing a function for the flexibility of the auditory system in terms of its ability to recognize sounds from a wide range of cues.

The ability to fall back on secondary cues provides a particular challenge for experimentalists. If one specific cue is preserved in an experimental setting, and listeners are able to perform the recognition task successfully, this does not mean that the same cue would be used if a rich set of cues was available, as would be expected in more naturalistic settings. More generally, if a search is targeted at a specific acoustic representation (e.g., the spectrogram or modulation spectrum) then usable features are likely to be found there whether or not these are the cues that the listener would normally use. Where there is doubt as to whether primary or secondary cues are being used, listeners could be slowed by a limited set of cues (see Sect. 3.4) (Delorme et al. 2000; Agus et al. 2012).

Nearly all of the methods discussed in this chapter, including traditional psychophysical methods, reaction-time data, animal physiology, and brain-imaging techniques, build on an ability to manipulate sounds, whether in terms of distorting them or synthesizing them with properties that are relevant to the theoretical question at hand. Each new technique, manipulation, or form of synthesis has led to breakthroughs in the understanding of timbre: magnetic tape in the 1960s; digital synthesis techniques in the 1970s (for an account of pioneering attempts in an analysis-synthesis exploration of timbre recognition, see Risset and Wessel 1999); vocoders in the 1990s; and more recently, optimization-based synthesis strategies that are built on cochlear models (e.g., McDermott and Simoncelli 2011). The surgical manipulation of sounds that is now available can be directed by strong theories to provide the next wave of insights into timbre recognition.

Acknowledgements DP was supported by the ANR grants ANR-10-LABX-0087 and ANR-10-IDEX- 0001-02, and by the European Research Council (ERC ADAM No. 295603).

Compliance with Ethics Requirements Trevor Agus declares that he has no conflict of interest.

Clara Suied declares that she has no conflict of interest.

Daniel Pressnitzer declares that he has no conflict of interest.

References

Agus TR, Thorpe SJ, Pressnitzer D (2010) Rapid formation of robust auditory memories: insights from noise. Neuron 66:610–618

Agus TR, Suied C, Thorpe SJ, Pressnitzer D (2012) Fast recognition of musical sounds based on timbre. J Acoust Soc America 131:4124–4133. https://doi.org/10.1121/1.3701865

Agus TR, Paquette S, Suied C et al (2017) Voice selectivity in the temporal voice area despite matched low-level acoustic cues. Sci Rep 7:11526. https://doi.org/10.1038/s41598-017-11684-1

Andrillon T, Kouider S, Agus T, Pressnitzer D (2015) Perceptual learning of acoustic noise generates memory-evoked potentials. Curr Biol 25. https://doi.org/10.1016/j.cub.2015.09.027

Backhaus VH (1932) Über die Bedeutung der Ausgleichsvorgänge in der Akustik. Z Tech Phys 13:31–46

Beauchamp JW (1975) Analysis and synthesis of cornet tones using nonlinear interharmonic relationships. J Aud Eng Soc 23:778–795

Belin P, Zatorre RJ, Lafaille P et al (2000) Voice-selective areas in human auditory cortex. Nature 403:309–312. https://doi.org/10.1038/35002078

Benade AH (1990) Fundamentals of musical acoustics. Dover Publications, New York

Berger KW (1964) Some factors in the recognition of timbre. J Acoust Soc Am 36:1888–1891. https://doi.org/10.1121/1.1919287

Bigand E, Delbé C, Gérard Y, Tillmann B (2011) Categorization of extremely brief auditory stimuli: domain-specific or domain-general processes? PLoS One 6:e27024. https://doi.org/10.1371/journal.pone.0027024

Brimijoin OW, Akeroyd MA, Tilbury E, Porr B (2013) The internal representation of vowel spectra investigated using behavioral response-triggered averaging. J Acoust Soc Am 133:EL118–EL122. https://doi.org/10.1121/1.4778264

Cavanagh P (2005) The artist as neuroscientist. Nature 434:301–307

Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118:887–906

Clark M, Luce D, Abrams R et al (1963) Preliminary experiments on the aural significance of parts of tones of orchestral instruments and on choral tones. J Aud Eng Soc 11:45–54

Coath M, Denham SL (2005) Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience. Biol Cybern 93:22–30. https://doi.org/10.1007/s00422-005-0560-4

Delorme A, Richard G, Fabre-Thorpe M (2000) Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. Vis Res 40:2187–2200. https://doi.org/10.1016/S0042-6989(00)00083-3

Elliott CA (1975) Attacks and releases as factors in instrument identification. J Res Mus Ed 23:35–40. https://doi.org/10.2307/3345201

Elliott TM, Hamilton LS, Theunissen FE (2013) Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. J Acoust Soc Am 133:389–404. https://doi.org/10.1121/1.4770244

Formisano E, De Martino F, Bonte M, Goebel R (2008) “Who” is saying “what”? Brain-based decoding of human voice and speech. Science 322:970–973

Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci 6:1216–1223

Giordano BL, McAdams S (2006) Material identification of real impact sounds: effects of size variation in steel, glass, wood, and plexiglass plates. J Acoust Soc Am 119:1171–1181. https://doi.org/10.1121/1.2149839

Giordano BL, McAdams S (2010) Sound source mechanics and musical timbre perception: evidence from previous studies. Music Percept 28:155–168

Gosselin F, Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. Vis Res 41:2261–2271. https://doi.org/10.1016/S0042-6989(01)00097-9

Gray GW (1942) Phonemic microtomy: the minimum duration of perceptible speech sounds. Commun Monogr 9:75–90

Grey JM (1977) Multidimensional perceptual scaling of musical timbres. J Acoust Soc Am 61:1270–1277

Grey JM, Moorer JA (1977) Perceptual evaluations of synthesised musical instrument tones. J Acoust Soc Am 62:454–462

Gygi B, Kidd GR, Watson CS (2004) Spectral-temporal factors in the identification of environmental sounds. J Acoust Soc Am 115:1252–1265

Gygi B, Kidd GR, Watson CS (2007) Similarity and categorization of environmental sounds. Percept Psychophys 69:839–855

Gygi B, Shafiro V (2011) The incongruency advantage for environmental sounds presented in natural auditory scenes. J Exp Psychol Hum Percept Perform 37:551–565. https://doi.org/10.1037/a0020671

Helmholtz H (1877) Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik, 4th edn. F. Vieweg und Sohn, Braunschweig. English edition: Helmholtz H (1954) On the sensations of tone as a physiological basis for the theory of music (trans: Ellis AJ), 2nd edn. Dover, New York

Hjortkjær J, McAdams S (2016) Spectral and temporal cues for perception of material and action categories in impacted sound sources. J Acoust Soc Am 140:409–420. https://doi.org/10.1121/1.4955181

Hjortkjær J, Kassuba T, Madsen KH et al (2018) Task-modulated cortical representations of natural sound source categories. Cereb Cortex 28:295–306. https://doi.org/10.1093/cercor/bhx263

Hromádka T, Zador AM (2009) Representations in auditory cortex. Curr Opin Neurobiol 19:430–433. https://doi.org/10.1016/S0959-4388(09)00096-8

Isnard V, Taffou M, Viaud-Delmon I, Suied C (2016) Auditory sketches: very sparse representations of sounds are still recognizable. PLoS One 11. https://doi.org/10.1371/journal.pone.0150313

Iverson P, Krumhansl CL (1993) Isolating the dynamic attributes of musical timbre. J Acoust Soc Am 94:2595–2603. https://doi.org/10.1121/1.407371

Kang H, Agus TR, Pressnitzer D (2017) Auditory memory for random time patterns. J Acoust Soc Am 142:2219–2232. https://doi.org/10.1121/1.5007730

Kell AJE, Yamins DLK, Shook EN et al (2018) A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98:630–644. https://doi.org/10.1016/j.neuron.2018.03.044

Kumar S, Bonnici HM, Teki S et al (2014) Representations of specific acoustic patterns in the auditory cortex and hippocampus. Proc R Soc B Biol Sci 281:20141000. https://doi.org/10.1098/rspb.2014.1000

Lakatos S, McAdams S, Causse R (1997) The representation of auditory source characteristics: simple geometric form. Percept Psychophys 59:1180–1190

Leaver AM, Rauschecker JP (2010) Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci 30(22):7604–7612. https://doi.org/10.1523/JNEUROSCI.0296-10.2010

Lewis JW, Wightman FL, Brefczynski JA et al (2004) Human brain regions involved in recognizing environmental sounds. Cereb Cortex 14:1008–1021. https://doi.org/10.1093/cercor/bhh061bhh061

Luo S, Zhu L, Althoefer K, Liu H (2017) Knock-knock: acoustic object recognition by using stacked denoising autoencoders. Neurocomputing 267:18–24. https://doi.org/10.1016/j.neucom.2017.03.014

McAdams S (1993) Recognition of sound sources and events. In: McAdams S, Bigand E (eds) Thinking in sound: the cognitive psychology of human audition. Oxford University Press, Oxford, pp 146–198

McAdams S, Winsberg S, Donnadieu S et al (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res 58:177–192. https://doi.org/10.1007/BF00419633

McAdams S, Beauchamp JW, Meneguzzi S (1999) Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. J Acoust Soc Am 105:882–897. https://doi.org/10.1121/1.426277

McDermott JH, Simoncelli EP (2011) Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron 71:926–940. https://doi.org/10.1016/j.neuron.2011.06.032

McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. Nat Neurosci 16:493–498. https://doi.org/10.1038/nn.3347

McPherson MJ, McDermott JH (2018) Diversity in pitch perception revealed by task dependence. Nat Hum Behav 2:52–66. https://doi.org/10.1038/s41562-017-0261-8

Newton MJ, Smith LS (2012) A neurally inspired musical instrument classification system based upon the sound onset. J Acoust Soc Am 131:4785–4798. https://doi.org/10.1121/1.4707535

Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88:1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Occelli F, Suied C, Pressnitzer D et al (2016) A neural substrate for rapid timbre recognition? Neural and behavioral discrimination of very brief acoustic vowels. Cereb Cortex 26:2483–2496. https://doi.org/10.1093/cercor/bhv071

Ogg M, Slevc LR, Idsardi WJ (2017) The time course of sound category identification : insights from acoustic features. J Acoust Soc Am 142:3459–3473

Patil K, Pressnitzer D, Shamma S, Elhilali M (2012) Music in our ears: the biological bases of musical timbre perception. PLoS Comput Biol 8(11):e1002759. https://doi.org/10.1371/journal.pcbi.1002759

Pressnitzer D, Agus T, Suied C (2015) Acoustic timbre recognition. In: Jaeger D., Jung R. (eds) Encyclopedia of computational neuroscience. Springer, New York, pp. 128–133

Risset J-C, Wessel DL (1999) Exploration of timbre by analysis and synthesis. In: The psychology of music, pp 113–169. https://doi.org/10.1016/B978-012213564-4/50006-8

Robinson K, Patterson RD (1995) The duration required to identify the instrument, the octave, or the pitch chroma of a musical note. Music Percept 13:1–15. https://doi.org/10.2307/40285682

Rosch E, Mervis CB (1975) Family resemblances: studies in the internal structure of categories. Cogn Psychol 7:573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Saldanha EL, Corso JF (1964) Timbre cues and the identification of musical instruments. J Acoustic Soc Am 36:2021–2026

Shannon RV, Zeng FG, Kamath V et al (1995) Speech recognition with primarily temporal cues. Science 270:303–304

Sidwell A, Summerfield Q (1985) The effect of enhanced spectral contrast on the internal representation of vowel-shaped noise. J Acoust Soc Am 78:495–506

Siedenburg K, Fujinaga I, McAdams S (2016) A comparison of approaches to timbre descriptors in music information retrieval and music psychology. J New Mus Res 45:27–41. https://doi.org/10.1080/09298215.2015.1132737

Siedenburg K, McAdams S (2017) Four distinctions for the auditory "wastebasket" of timbre. Front Psychol 8:1747. https://doi.org/10.3389/fpsyg.2017.01747

Suied C, Agus TR, Thorpe SJ, Pressnitzer D (2013a) Processing of short auditory stimuli: the rapid audio sequential presentation paradigm (RASP). In: Basic aspects of hearing. Springer, New York, pp 443–451

Suied C, Drémeau A, Pressnitzer D, Daudet L (2013b) Auditory sketches: sparse representations of sounds based on perceptual models. In: Aramaki M, Barthet M, Kronland-Martinet R, Ystad S (eds) From sounds to music and emotions. CMMR 2012. Lecture Notes in Computer Science. Springer, Berlin/Heidelberg

Suied C, Agus TR, Thorpe SJ et al (2014) Auditory gist: recognition of very short sounds from timbre cues. J Acoust Soc Am 135:1380–1391. https://doi.org/10.1121/1.4863659

Thoret E, Depalle P, McAdams S (2016) Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments. J Acoust Soc Am 140:EL478–EL483. https://doi.org/10.1121/1.4971204

Thoret E, Depalle P, McAdams S (2017) Perceptually salient regions of the modulation power spectrum for musical instrument identification. Front Psychol 8:587. https://doi.org/10.3389/fpsyg.2017.00587

Tomasino B, Canderan C, Marin D et al (2015) Identifying environmental sounds: a multimodal mapping study. Front Hum Neurosci 9:567. https://doi.org/10.3389/fnhum.2015.00567

Venezia JH, Hickok G, Richards VM (2016) Auditory "bubbles": efficient classification of the spectrotemporal modulations essential for speech intelligibility. J Acoust Soc Am 140:1072–1088. https://doi.org/10.1121/1.4960544

Viemeister NF, Wakefield GH (1991) Temporal integration and multiple looks. J Acoust Soc Am 90:858–865

Warren WH, Verbrugge RR (1984) Auditory perception of breaking and bouncing events: a case study in ecological acoustics. J Exp Psychol Hum Percept Perform 10:704–712. https://doi.org/10.1037/0096-1523.10.5.704

Wedin L, Goude G (1972) Dimension analysis of the perception of instrumental timbre. Scand J Psychol 13:228–240. https://doi.org/10.1111/j.1467-9450.1972.tb00071.x

Wun S, Horner A, Wu B (2014) Effect of spectral centroid manipulation on discrimination and identification of instrument timbres. J Aud Eng Soc 62:575–583. https://doi.org/10.17743/jaes.2014.0035

Yonelinas AP (2002) The nature of recollection and familiarity: a review of 30 years of research. J Mem Lang 46:441–517. https://doi.org/10.1006/jmla.2002.2864

Chapter 4
Memory for Timbre

Kai Siedenburg and Daniel Müllensiefen

Abstract Memory is a cognitive faculty that is of fundamental importance for human communication in speech and music. How humans retain and reproduce sequences of words and pitches has been studied extensively in the cognitive literature. However, the ability to retain timbre information in memory remains less well understood. Recent years have nonetheless witnessed an upsurge of interest in the study of timbre-related memory processes in experimental psychology and music cognition. This chapter provides the first systematic review of these developments. Following an outline of basic memory concepts, three questions are addressed. First, what are the memory processes that govern the ways in which the timbres of sound sequences are recognized? Predominantly focusing on data from short-term recognition experiments, this review addresses aspects of capacity and similarity, sequential structures, and maintenance processes. Second, is there interference of timbre with other attributes in auditory memory? In other words, how specific are memory systems for timbre and to what degree are they separate from memory systems for pitch and verbal information. Third, do vocal sounds and the sounds from familiar sources possess a special status in auditory memory and, if so, what could be the underlying mechanisms? The chapter concludes by proposing five basic principles of memory for timbre and a discussion of promising avenues for future research.

Keywords Acoustic similarity · Auditory memory · Melodic memory · Memory maintenance · Pitch interference · Sound-source familiarity · Verbal memory · Voice superiority · Working memory

K. Siedenburg (✉)
Department of Medical Physics and Acoustics, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
e-mail: kai.siedenburg@uni-oldenburg.de

D. Müllensiefen
Department of Psychology, Goldsmiths, University of London, London, United Kingdom
e-mail: d.mullensiefen@gold.ac.uk

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_4

4.1 Introduction

Memory, the capability to explicitly or implicitly remember past experiences, is one of the most extraordinary and mysterious abilities of the mind. Memory defines human perception, cognition, and identity. Speech and music, both fundamental to human nature and culture, are based on short- and long-term memory for acoustic patterns. Memories exist for many, but not all, experienced events: Think about which clothes you wore on an important day of your life versus which ones you wore last Wednesday (unless Wednesday was important). Not all aspects of perceptual experience are memorized equally well: Think about whether the first notes from a song you like go up or down versus the exact pitch height of the melody's first note.

While assessing memory for pitch patterns, tone sequences, and melodies has a long tradition in auditory psychology (e.g., Deutsch 1970; Müllensiefen and Halpern 2014), there are considerably fewer publications on memory for timbre. Hence, does memory for timbre exist at all? Are humans able to remember timbre information, such as the quality of an unfamiliar voice or the sonority of a particular sound sequence from a music track, over short and long time spans? Or is timbre an attribute of auditory experience that is reserved for being experienced in the moment? Only 10 years ago, research did not have proper empirical ground to answer these questions. In fact, it is important to note that timbre has not been considered critical for memory and cognition for a long time.

One reason for the lack of research on memory for timbre is the fact that speech and music have most commonly been considered within an information processing framework (e.g., Simon 1978), whereby the communicative message is conceptualized as sequences of phonemic or pitch categories that are independent from the properties of the carrier medium, which includes the sounds' vocal or instrumental timbre. Influential models of human memory (Atkinson and Shiffrin 1968) presumed that aspects of sensory information could be transformed into cognitive information and short-term or long-term memory using symbolic recoding. Any sensory information that could not be recoded was assumed to be lost from the sensory (echoic) memory store (Darwin et al. 1972). A second reason for the lack of research on timbre memory might be rooted in the fact that classic music-theoretical approaches—traditionally a driving force behind much music cognition research (Meyer 1956; Lerdahl and Jackendoff 1983)—focus on pitch and duration and their derived musical parameters harmony and rhythm but do not cover timbre as a primary musical parameter. Thirdly, the relative scarcity of empirical evidence for complex cognitive processes related to timbre, such as effects of auditory context or musical experience, may have had additional impact. Overall, this situation may have created the false impression that timbre is an auditory surface feature that is not essential to the cognitive architecture of human audition. Fortunately, this situation is beginning to change and many researchers in experimental psychology and music cognition have started to address timbre-related questions.

As summarized in this chapter, the effects of auditory context and long-term auditory experience with timbre have been demonstrated (both at a behavioral and neural level), the role of voice timbre in speech perception has become subject to experimental scrutiny, and the effects of vocal timbre on verbal memory have been known for a longer time. Clearly, timbre is becoming a rich and exciting topic for auditory cognition research, and memory obviously plays an important role in this development. Note that of the forty-five or so empirical studies on memory for timbre, more than thirty-five have been published between 2008 and 2018. This chapter provides the first systematic review of this emerging field and thereby highlights the fact that memory for timbre is a highly relevant concept in auditory cognition.

Section 4.2 describes general concepts from memory research, in particular, with regards to auditory memory systems for short-term and long-term storage, the granularity of auditory memory, and models of short-term memory. Regarding memory for timbre, four research themes stand out and provide a structure for subsequent sections. The first research theme comprises many studies that scrutinize the structure of short-term memory for timbre and have started to propose cognitive mechanisms that might be implicated. In Sect. 4.3, the presumed capacity limits of short-term memory for timbre will be discussed with a particular focus on the role of perceptual similarity and chunking.

The second theme concerns the active maintenance and imagery of timbre, which is addressed in Sect. 4.4. The tenet of this section is that memory for timbre involves elements of attentional control, which recreate facets of auditory experience.

The third theme focuses on the growing body of work that is demonstrating interference from auditory attributes on primary memory contents. For instance, variability in a task-irrelevant attribute, such as timbre, strongly impairs performance in a melodic memory task wherein the primary content (i.e., melodic structure) is conceptually independent of timbre. These findings are described in Sect. 4.5, which discusses the status of memory representations for timbre: Are they stored separately from other auditory attributes such as pitch or verbal information?

The fourth theme, discussed in Sect. 4.6, focuses on the role of sound source familiarity in memory for timbre and effects of voice superiority. Several studies have reported processing advantages for vocal timbre over timbres from nonvocal musical instruments. This finding resonates with the assumption that human listeners are specialists in voice timbre processing. To synthesize our discussion, five principles of memory for timbre are proposed that address some of the underlying cognitive processes. For a more general discussion of auditory memory, please see Demany and Semal (2007). A treatment of sound (source) recognition is included in Chap. 3 (Agus, Suied, and Pressnitzer) of this volume.

4.2 Auditory Memory Concepts

Memory is an overwhelmingly broad notion that plays a central role in almost every aspect of human cognition. At its core is the retention over time of experience-dependent internal representations and the capacity to reactivate such representations (Dudai 2007). Representations of sensory information and cognitive states thus are starting points for the formation of memories. But it is the temporal trajectory of these representations that defines memory and makes it such a rich and complex research topic.

4.2.1 Stores and Processes

An elementary conceptual distinction regarding the structure of human memory concerns the differences between the short-term and long-term memory systems. William James (1890/2004) already thought of primary (conscious, short-lived) and secondary (unconscious, long-lived) memory as independent entities. A more fine-grained distinction became the core of the classic *multistore* or *modal model*, most prominently elaborated by Atkinson and Shiffrin (1968). It posits three types of stores, namely a sensory register, a short-term memory (STM) store, and a long-term memory (LTM) store. According to Atkinson and Shiffrin (1968), sensory information is subject to modality-specific, pre-attentive storage of fast decay (within 2 s) unless there is a subject-controlled scan via selective attention, which recodes and transfers portions of the register to the short-term store. This store is thought to retain a categorical, modality-independent code where traces decay within time spans of less than 30 s. Their life spans can be lengthened by active rehearsal, which lends them more time to be transferred to the long-term store.

To refine this classic picture, Cowan (1984, 2015) proposed a taxonomy for nonverbal auditory memory that emphasized similarities with visual memory. In vision, one can find a seemingly clear structural divide between an automatic sensory storage of almost unlimited capacity and fast decay (< 200 ms)—*iconic memory*—and a more long-lived, attention-dependent, short-term memory system of constrained capacity. Cowan's *short auditory store* is hypothesized to be experienced as sensation or sensory afterimage (i.e., is distinct from the sensory type of memory required to integrate and bind perceptual features, such as loudness or amplitude modulations, over tenths of seconds). The short auditory store contains not-yet-analyzed, pre-categorical content that decays within 200–300 ms. The *long auditory store* is experienced as (short-term) memory, contains partially analyzed or categorized content, and is supposed to decay within 2–20 s. Due to the structural similarity of the long store and categorical STM (Atkinson and Shiffrin 1968) with regard to decay rates and capacity, Cowan considered the long auditory store to be a special case of STM. Contrary to the classic multistore models that assume that STM operates on verbal items, Cowan's proposal implies that STM may also operate on sensory representations.

Although Cowan's distinction between a short and automatic versus a long and consciously controlled form of auditory memory may have intuitive appeal due to its analogy to vision, recent data suggest that it is hard to find clear-cut boundaries. Several studies have highlighted difficulties in estimating the exact duration of the shorter type of auditory memory. More specifically, testing the discrimination of frequency shifts within nonharmonic tone complexes, Demany et al. (2008) observed a gradual decay in performance for increasing retention times, which is not comparable to the steep decline that is characteristic of iconic memory in vision. Importantly, there was no clear evidence for differential memory capacity (i.e., a short store of high capacity and a long store of low capacity) within the 2 s range of retention times tested. Demany et al. (2010) explicitly compared visual and auditory change detection. Whereas visual memory fidelity appeared to decay quickly and substantially within 200 ms, confirming the classical view on iconic memory, there was no such sign for auditory memory, which persisted throughout retention times of 500 ms at much lower decay rates. This finding indicates that auditory change detection may operate on much longer time scales than visual iconic memory. As a theoretical explanation, Demany et al. suggest frequency shift detectors as a cognitive mechanism that tracks spectral changes of stimuli. These detectors were shown to gradually lose their tuning specificity when inter-stimulus intervals increase (Demany et al. 2009). But the observed differences in tuning specificity were gradual rather than showing clear-cut boundaries.

Rejecting the idea of clear separations between hypothetical memory stores resonates with the *proceduralist approach* to memory (Crowder 1993; Jonides et al. 2008). Instead of conceptualizing memory as a separate cognitive system, implemented by a multitude of interacting modules (e.g., sensory, STM, and LTM), the unitary or proceduralist approach understands memory as an emergent property of the ways in which mental processes operate on perceptual representations or cognitive states. As noted by Craik and Lockhart (1972), “It is perfectly possible to draw a box around early analyses and call it sensory memory and a box around intermediate analyses called short-term memory, but that procedure both oversimplifies matters and evades the more significant issues” (p. 675). A classical illustration of the idea of memory being a byproduct of perceptual processing is given by the *levels of processing effect* (Craik and Lockhart 1972): If experimental participants’ attention in an encoding phase is drawn toward “deep” semantic features of words (as in a semantic categorization task), recall is better than if participants judge “shallow” perceptual features of the stimuli (as in phonemic categorization). Contemporary neuroimaging studies support unitary views of memory in the sense that, in general, the same neural ensembles are found to be responsible for perceptual processing and memory storage (D’Esposito and Postle 2015).

Note that even if one does not believe in the existence of dedicated short-term and long-term memory systems, the notions of STM and LTM may be used as referents to memory function over short or long time intervals. This agnostic usage acknowledges that there may be different time scales of memory persistence but does not presuppose any particular stores or cognitive mechanisms.

4.2.2 Granularity of Auditory Memory

Another line of research has raised the question of how fine-grained auditory memory representations can be. In other words, what is the smallest detail of a sound that can be remembered? Using noise waveforms that are completely identical according to macroscopic auditory features, such as spectral and temporal envelope, Kaernbach (2004) showed that repetitions of noise segments could be well detected up to at least 10 s of segment length; single, seamless repetitions of noise waveforms were detected with above-chance accuracy up to 2 s. Agus et al. (2010) even demonstrated that there is a form of long-term persistence for features of noise waveforms (also see Agus, Suied, and Pressnitzer, Chap. 3). When requiring listeners to detect repetitions of noise segments, recurring noise stimuli featured far superior hit rates compared to novel noise waveforms. Notably, subjects were not aware that segments reoccurred and must have implicitly picked up idiosyncratic features of the presented noise tokens. This demonstrates that there is implicit, nondeclarative long-term auditory memory even for small sensory details. This memory process appears to be fully automatic: Andrillon et al. (2017) even demonstrated that noise snippets were memorized during rapid-eye-movement sleep.

What is the relation between this detailed form of memory and the formation of general auditory categories? McDermott et al. (2013) had listeners discriminate different classes of resynthesized environmental textures (e.g., rain versus waves) and exemplars of textures (e.g., one type of rain versus another). Texture category discrimination performance gradually increased with excerpt length (40–2500 ms) but, curiously, the discrimination of exemplars within categories gradually worsened. This was interpreted as an indication that summary statistics underlie the representation of sound textures: Representations of two exemplars from the same category converge with increasing excerpt length because averaging over increased lengths removes idiosyncratic sound features. In sum, this implies that humans can possess fine-grained memories of auditory events (Agus et al. 2010), but the recognition of sound (texture) categories likely relies on robust summary statistics that are less affected by idiosyncratic details (McDermott et al. 2013).

4.2.3 Capacity Limits in Short-Term Memory

A common assumption in studies of human short-term memory is its limited capacity. The famous conjecture by Miller (1956) states that people can retain 7±2 independent chunks of information in immediate memory. This idea has been of enormous impact in cognitive (and popular) science. Miller's core idea was that the informational bottleneck of short-term memory does not strictly depend on the number of items, but that there is a general limit on the number of *independent chunks* of information in short-term memory. The concept of item and chunk are distinct because sequences of items may be recoded into fewer chunks. More technically, a chunk can be defined as a "collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use" (Cowan 2001, p. 89). For example, sequences of letters, such as IRSCIAFBI, are far easier to memorize when remembered as chunks IRS CIA FBI (familiar US federal agencies) than as raw item-by-item successions (Cowan 2008).

Presenting a contemporary revision of Miller's original hypothesis, Cowan (2001) reviewed empirical evidence across a wide range of domains such as verbal, visual, and auditory memory. Specifically, Cowan argued that the capacity limit of short-term memory (STM) is only about 4±1 chunks if the involvement of other factors, such as long-term memory (LTM) and active rehearsal, is limited. The above example illustrates this proposal because long-term memory enables participants to form chunks such as IRS, CIA, and FBI. The role of active rehearsal, classically considered as vocal or subvocal (i.e., silent) repetition of the stimuli in verbal memory research (Baddeley 2012), would be to actively maintain the memory trace.

Despite its considerable influence, Cowan's 4±1 proposal has received harsh criticism from the very beginning (see the peer commentaries in Cowan 2001). An alternative framework that has gained momentum in visual memory research replaces the idea of magical numbers in STM (7±2 or 4±1) by *resource-based models* (Ma et al. 2014). These models of short-term memory assume limited resources in terms of the representational space or medium shared by items but not a limit to the exact number of items that can be maintained. Stimulus representations are considered to be corrupted by noise. The level of the noise increases with more items to be held in memory because items interfere with each other in their representational space. In other words, resource models assume that short-term memory is fundamentally limited in the quality, rather than the quantity, of information. These assumptions imply an increased probability of memory lapses in situations when items are perceptually similar.

Transferring the concept of capacity limits or a capacity-similarity tradeoff to timbre entails the question of what constitutes the basic unit to be memorized, that is, the item. In the study of verbal memory, individual words naturally qualify as items because language is composed of strings of words. However, there are many other domains for which the situation is not as clear. As Ma et al. (2014) noted with regards to vision, "An 'item' is often relatively easy to define in laboratory experiments, but this is not necessarily the case in real scenes. In an image of a bike, for example, is the entire bike the item, or are its wheels or its spokes items?" Similar complications may be in place for auditory memory beyond speech. In the context of polyphonic music, there can be plenty of timbral contrast that arises in short time spans from the sounds of various instruments. But it is not intuitively clear what constitutes the unit of the item in this case: individual tones, fused auditory events, or segments of auditory streams? In analogy to the existing verbal memory research, many studies of STM for musical timbre (see Sect. 4.3) use sequences of individual tones that differ by timbre, for instance, with sounds from diff erent orchestral instruments changing on a note-by-note basis. Although this operationalization may be seen as a plausible perceptual model for an orchestration technique, such as *Klangfarbenmelodie* (i.e., timbre melodies; Siedenburg and McAdams 2018; McAdams, Chap. 8) or percussion music (Siedenburg et al. 2016), it does not seem to be an appropriate model for many other types of music, for which this type of strong timbral contrast on a note-to-note basis represents a rare exception.

4.3 Factors in Short-Term Recognition

4.3.1 Memory Capacity and Similarity

Similarity effects are a hallmark of verbal and visual STM (Baddeley 2012; Ma et al. 2014). Despite being perceptually discriminable, similar items are more frequently confused in memory compared to dissimilar ones. With regards to memory for timbre, however, research is only beginning to account for effects of perceptual similarity relations.

Starr and Pitt (1997) used an interpolated tone paradigm (cf., Deutsch 1970) that required participants to match a standard and a comparison stimulus, separated by a 5 s interval with intervening distractor tones. Their first experiment demonstrated an effect of timbre similarity: The more similar in brightness the interfering tones were to the target tone, the more detrimental was their effect on retention in memory. Visscher et al. (2007) tested auditory short-term recognition in an item recognition experiment using auditory ripple stimuli (i.e., amplitude-modulated sinusoid complexes). They observed that two independent factors caused decreases in false alarm rates on a trial-by-trial basis: (a) increases of the mean dissimilarity of the probe sound to the sequence and (b) increases of the perceptual homogeneity of the sounds in the sequence, that is, the average similarity between the sounds in the sequence.

In one of the first studies, Golubock and Janata (2013) set out to measure capacity limits of short-term memory for timbre. They used an item recognition task with synthetic sounds differing by timbre (constituting the items). They synthesized sounds that varied along the dimensions of spectral centroid, attack time, and spectral flux, the discriminability of which was ensured via separate just-noticeable-difference measurements. Sequences of 2–6 tones that differed in timbre were presented, but the tones were of constant pitch and loudness. Each sequence was followed by a silent retention interval of 1–6 s, and then a single probe tone was presented for which participants had to judge whether it was part of the sequence or not. The authors observed memory capacities at around K = 1.5 items, estimated according to the formula

$$K = (\text{hit rate} + \text{correct rejection rate} - 1) * N,$$

where N denotes the number of items in the test sequence. Capacities significantly decreased with increasing sizes of the retention intervals, with K = 1.7 for 1 s and K = 1.3 for 6 s.

The large difference between the capacity estimate of an average of 1.5 timbre items from Golubock and Janata (2013) and the supposedly universal estimate of 3–5 items according to Cowan (2001) seems striking. Why should memory for timbre be so much worse? Notably, the sounds in Golubock and Janata's first experiment only varied along three timbral dimensions.

A second experiment used a more heterogeneous set of sounds from a commercial keyboard synthesizer and measured a significantly greater capacity of around 1.7 items. Figure 4.1 displays hit and correct rejection rates averaged across reten-

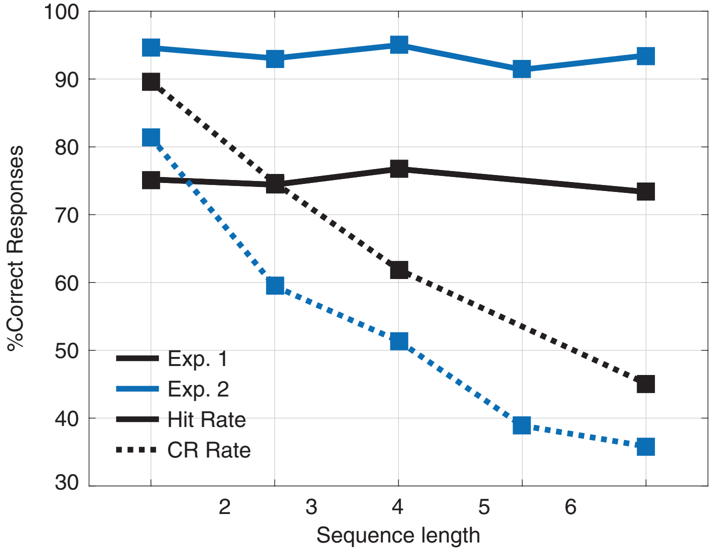

Fig. 4.1 Accuracy (percentage of correct responses) as a function of the sequence length in the two-item recognition experiments. Experiment 1 used abstract synthetic sounds; experiment 2 used sounds selected from a commercial sound sampler. Hits correspond to correct identification of match trials, correct rejections (*CR*) to correct identification of nonmatch trials. (Adapted from Table 1 in Golubock and Janata 2013; used with permission from the American Psychological Association)

tion intervals from the memory experiments in their study. Hit rates are higher in experiment 2 compared to experiment 1, but the false alarm rates of experiment 2 also exceed those of experiment 1. However, no trial-by-trial analyses of these data were conducted, and it remains unclear whether the increase in capacity in the second experiment was primarily caused by a global increase in the timbral homogeneity of sounds or by greater probe list dissimilarities.

Using an item recognition task, Siedenburg and McAdams (2017) observed significant correlations between participants' response choices (i.e., whether they recognized a probe sound as match or nonmatch) and the mean perceptual dissimilarity from the probe to the tones in the sequence. However, no significant correlation between timbral homogeneity and response choices was observed.

Siedenburg and McAdams (2018) further evaluated the role of similarity in a serial recognition task. They had participants indicate whether the order of the timbres of two subsequently presented sound sequences was identical or not. In the non-identical case, two sounds were swapped. A correlation analysis showed that the timbral dissimilarity of swapped items (TDS) was a good predictor of response choice in serial recognition and predicted around 90% of the variance of response choices throughout four experiments. This study also tested for the role of sequence homogeneity but did not find a consistent effect: Homogeneity and response choice were significantly correlated in only one out of four experiments. Moreover,

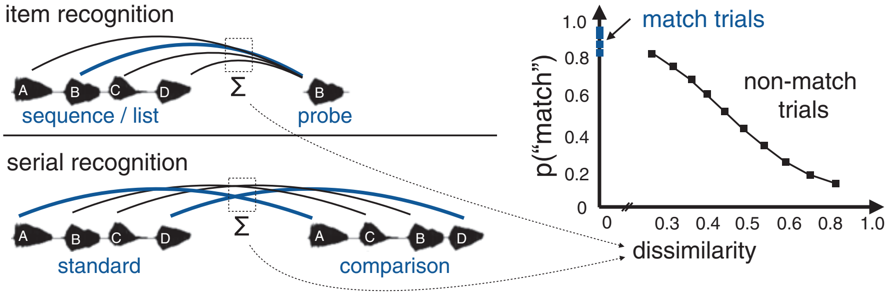

Fig. 4.2 Schematic depiction of the relationship between response choice (probability of "match" responses) and timbre dissimilarity. For item recognition tasks, the hypothetical dissimilarity measure corresponds to the sums of dissimilarities (Σ) of the probe item to all the items in the sequence (indicated by *connecting lines*). The *blue line* indicates a match and, hence, zero dissimilarity. For serial recognition tasks, the dissimilarity measure could be derived from the sum of the item-wise dissimilarities, resulting in the dissimilarity of the two items that were swapped (here: items *C* and *B*). Dissimilarity is normalized between 0 and 1

stepwise regression analysis failed to include homogeneity as a predictor of response choices in any experiment, indicating that a parsimonious account would not consider homogeneity as a crucial factor for timbre recognition. Figure 4.2 provides a schematic visualization of the described relation between response choice and both the probesequence dissimilarity in item recognition and the timbral dissimilarity of the swap in serial recognition.

Taken together, the strong effects of similarity (Siedenburg and McAdams 2018) and the wide range of estimates for timbre STM capacity (that differ clearly from STM capacity estimates for other auditory material; Golubock and Janata 2013) indicate that fixed-slot models of STM capacity may not be suitable as a model of STM for timbre. On the contrary, resource-based approaches that assume limited representational resources, and thus take into account similarity relations from the very beginning, appear to be better suited for the data from timbre experiments, although no formal model evaluation has been conducted yet. This is in line with the observed trade-off between the number of items that can be maintained in short-term memory and their timbral similarity.

4.3.2 Sequential Chunking

As already mentioned in Sect. 4.2.3, many memory studies try to avoid sequences with an explicit sequential structure. In order to measure memory proper, the rationale is that sequences should not explicitly allow for chunking made possible through grouping or repetition (Cowan 2001). At the same time, it is likely that affordances for sequential processing are important ecological factors in memory for timbre.

Siedenburg et al. (2016) considered the case of timbral sequencing as part of the tabla drumming tradition from North India. The tabla is a pair of hand drums with an extremely rich timbral repertoire and is considered the most important percussion instrument in North Indian classical music (Saxena 2008). Tabla music exhibits intricate serial patterns with hierarchical dependencies, for instance, through the nested repetition of groups of sounds. The centuries old tradition of tabla is taught as part of an oral tradition. Compositions are learned via the memorization of sequences of bols, that is, solfège-like vocalizations associated with drum strokes. In tabla solo performances, the verbal recitation of the composition oftentimes precedes the actual drumming. Furthermore, North Indian classical music is unfamiliar to most (but not all) western listeners and hence is well-suited for exploration of the effects of long-term memory on sound sequence recognition.

The experiment compared the recognition of tabla sequences between a group of tabla students and a group of western musicians unfamiliar with tabla music. As depicted in Fig. 4.3, four distinct sequencing conditions were used in the experiment: (1) idiomatic tabla sequences, (2) reversed sequences, (3) sequences of random order, and (4) sequences of random order and randomly drawn items without replacement. In the serial order recognition experiment, participants indicated whether the sounds in two consecutively played sequences were presented in the same order.

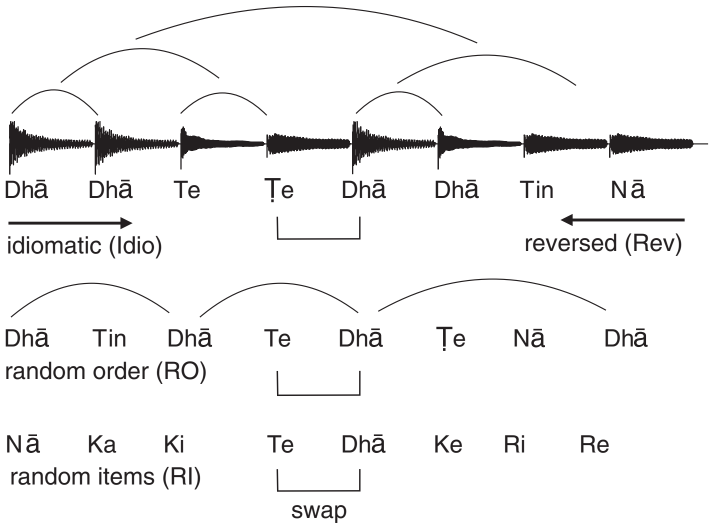

Fig. 4.3 Examples of the four sequencing conditions: an *idiomatic* sequence of bols (*Dha, Te, Tin, Na*) and the corresponding *reversed*, *random order*, and *random items* (adding *Ke, Ri, Re*) conditions (drawn without replacement). Note that in the idiomatic, reversed, and random order condition, there are items that occur multiple times in the sequence. (From Siedenburg et al. 2016; used with permission of the American Psychological Association)

The results showed a very strong effect of sequential structure: Idiomatic sequences of tabla strokes and their reversed versions were recognized best, followed by their counterparts with randomly shuffled order, followed by fully random sequences without repetitions of items. The latter effect indicated a facilitation of chunking due to the repetition of items. Because serial-order recognition was tested, it could be concluded that the advantage of redundancy primarily goes back to chunking and not a reduced load in terms of item identity. The advantage of reversed sequences over randomly shuffled ones was suspected to be related to the hierarchical structure inherent in the idiomatic sequences or their reversed versions. The reversed versions not only contained item repetitions, but repeating subsequences of items, such that sequences could be encoded hierarchically. Notably, effects of familiarity with idiomatic sequences (comparing tabla students versus naıve controls) only occurred for the vocal sounds but not for the drum sounds. This result indicates that vocal sounds are particularly well suited for chunking via long-term associations. Participants who are familiar with tabla can simply represent idiomatic sequences of bols (tabla words) via one item and hence have a significant mnemonic advantage over naıve participants. However, memory for instrumental sounds did not follow the same pattern, which may indicate that familiarity-based chunking is particularly effective for vocal sounds for which humans have a natural proficiency for combining basic building blocks in endless ways (e.g., Hagoort and Indefrey 2014).

An example of long-term recognition of timbre sequences was provided by Tillmann and McAdams (2004) who adopted the sequence-learning paradigm made famous by Saffran et al. (1999). Their results indicated that memory for timbre sequences is strongly affected by grouping cues provided by perceptual dissimilarity relations between subsequent tone pairs in the sequences (for more information, see McAdams, Chap. 8).

From a general perspective, these results indicate that auditory sequences can be stored much more efficiently if chunked in appropriate ways. Chunking could make memory for sequences more robust by structuring the memory trace along a hierarchy of time scales that is provided by grouping cues (if this sounds abstract, think about how to memorize, ABCXYZABCQ). This perspective allows us to explain effects in both short-term (Siedenburg et al. 2016) and long-term recognition (Tillmann and McAdams 2004).

4.4 Active Maintenance and Imagery of Timbre

In this section, it is argued that memory for timbre is not a fully automatic process that is solely based on persistence of passive information. Timbre representations can be consciously refreshed in working memory and recreated from long-term memory.

4.4.1 Maintenance in Working Memory

The key property that distinguishes the concept of working memory (WM) from that of short-term memory is the role of active manipulation and maintenance of the memory contents (although both terms are often used interchangeably). In contrast to the presumably passive and automatic process of auditory short-term memory, WM is usually defined as an active form of memory that, as a whole, underpins a range of important cognitive faculties such as problem solving and action control. The active nature of verbal WM becomes apparent when thinking of how phone numbers, street names, or vocabulary words in foreign language classes are commonly memorized. People tend to vocalize, openly or covertly, in order to retain verbal information in mind. This observation was captured by Baddeley's influential multicomponent model of working memory, which described verbal WM as governed by a phonological storage buffer and a rehearsal mechanism, overall giving rise to the *phonological loop* (Baddeley 2012). The memory trace in the buffer would decay gradually but could be refreshed by (sub)vocal rehearsal in order to be kept in the loop. In other words, the original auditory event undergoes a form of recoding into a sensorimotor code that allows conscious rehearsal.

Because of their success in explaining verbal working memory, the concept of the phonological loop has also influenced nonverbal auditory memory research and research into melodic memory in particular (Berz 1995; Schulze and Koelsch 2012). More specific to our concerns is the question of whether nonverbal auditory working memory and STM for timbre are subject to similar active maintenance processes. In other words, in which sense is short-term memory for timbre *working*?

Nees et al. (2017) tested whether melodic short-term recognition is supported by active rehearsal or by attention-based processes. Using a sequence matching task, participants listened to two melodies separated by an 8 s retention interval and judged the melodies as identical or non-identical. As is common in verbal WM research, a dual task paradigm was used. The basic assumption is that if a secondary task severely impairs the accuracy in the target task, the latter can be assumed to rely on similar cognitive processes and resources. Nees et al. (2017) used four secondary task conditions. An articulatory suppression (AS) condition required participants to read out loud solved math problems that were presented visually (e.g., 2 + 3 = 5). In an attentional refreshing suppression (ARS) condition, participants silently read math problems presented on a screen and needed to type the correct response on a computer keyboard. A third condition combined both articulatory and attentional refreshing suppression by having participants read aloud the math problem and provide the response orally (AS+ARS). A silent condition without suppression served as a baseline. Notably, the authors found that performance did not differ between the control and the ARS condition, but both the AS and AS+ARS conditions yielded a marked decline of sensitivity. These results indicate that melodic short-term memory is supported by subvocal rehearsal and not by attentional refreshing, suggesting strong structural similarities to verbal memory. As described in the following, the clarity of these findings for melody recognition by Nees et al. (2017) differs from the situation that we find for timbre.

Three distinct mechanisms for the maintenance of timbre in WM appear to be possible a priori. First, timbre recognition could be a passive process, which would imply that maintenance in fact does not play a strong role. The retention of timbre would instead primarily rely on the persistence of the sensory memory trace. Second, participants could attach labels to timbres (e.g., piano-violin-weird voice) and subsequently rehearse the verbal labels. This would constitute a verbal surrogate of memory for timbre. Third, listeners could allocate attention to the auditory memory trace and mentally replay timbre representations in their minds, a process that has been called *attentional refreshing* (Camos et al. 2009).

Several studies have gathered data that have implications for deciding on the plausibility of the mechanisms. McKeown et al. (2011) had three participants discriminate small changes in the spectral distribution of tones and showed that sensitivity was above chance even for extended retention intervals of 5–30 s. This effect was robust to an articulatory suppression task in which participants were required to read aloud during the retention time. These results were interpreted as evidence for a type of sensory persistence that is neither based on verbal labeling nor due to attentional refreshing. Schulze and Tillmann (2013) compared the serial recognition of timbres, pitches, and words in various experimental variants, using sampled acoustical-instrument tones and spoken pseudowords. They found that the retention of timbre, contrary to that of pitches and words, did not suffer from concurrent articulatory suppression, speaking against the involvement of labeling. In line with McKeown et al. (2011), they concluded that STM for timbre is structured differently than working memory for words or pitches and is unlikely to be facilitated by verbal labeling and (sub)vocal rehearsal. Nonetheless, their results did not rule out the possibility of attentional refreshing.

On the other hand, there are studies that have underlined the necessity of attentional refreshing for maintaining timbre information in memory. Soemer and Saito (2015) observed that short-term item recognition of timbre was only inconsistently disrupted by articulatory suppression but was more strongly impaired by a concurrent auditory imagery task. The authors interpreted these results as evidence that memory for timbre can be an active, re-enacting process that relies on the support of attentional resources. Siedenburg and McAdams (2017) more directly compared the effect of articulatory suppression with a suppression condition that captured listeners' visual attention. They used an item recognition task with familiar and unfamiliar sounds that were controlled for their timbral similarity relations. Three different suppression tasks filled the 6 s retention interval between the sound sequence and the probe sound. Participants either waited in silence, counted out loud (articulatory suppression), or detected identical exemplars in sequences of black and white grids (visual suppression). Results showed a clear advantage for familiar sounds that persisted throughout all experimental conditions. Surprisingly, there was no difference between articulatory and visual suppression, neither for familiar nor for unfamiliar sounds. However, both types of suppression affected timbre memory negatively compared to the silence condition.

Considering these empirical results from Siedenburg and McAdams (2017), multiple reasons speak for attentional refreshing as an important maintenance strategy for timbre. Firstly, verbal labeling was unlikely to act as a dominant maintenance strategy for timbre: Performance on unfamiliar sounds that were difficult to label was impaired under both articulatory and visual suppression. It seems much more plausible that the detrimental effect of articulatory suppression was due to interference with the auditory trace. Secondly, the plausibility of passive sensory storage without any active maintenance was ruled out by the detrimental effect of visual suppression, which should not interfere if auditory and visual WM are fully separated. Finally, it could be assumed that attentional refreshing was moderately disrupted by both types of suppression because the visual distractor task reduced attentional resources that refreshing relies on, and articulatory suppression interfered with the auditory trace that is subject to refreshing (beyond rather minor attentional requirements). Overall, these reasons indicated that attentional refreshing was the most likely candidate for active maintenance of timbre in the experiments by Siedenburg and McAdams (2017). The results by McKeown et al. (2011), to the contrary, indicated that neither verbal labeling and rehearsal nor attentional refreshing was necessary for successful timbre recognition. Taken together, this finding suggests that attentional refreshing is likely a sufficient, but not a necessary, condition of WM for timbre.

4.4.2 *Mental Imagery of Timbre*

Beyond the realm of maintaining information in short-term memory, research has also provided evidence for the feasibility of a closely related mental faculty: imagery for timbre. Whereas attentional refreshing was understood as a sort of attention-driven

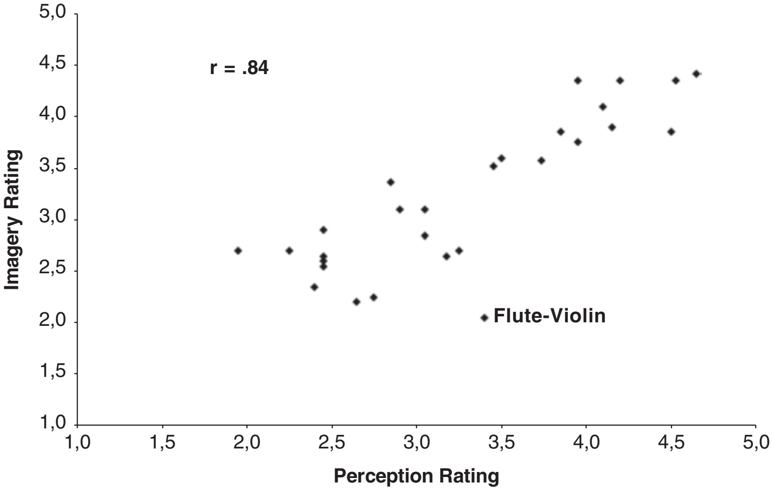

Fig. 4.4 Scatterplot of mean similarity ratings for each instrument pair in the perception and imagery conditions. Correlation coefficient, r = 0.84. (From Halpern et al. 2004; used with permission from Elsevier)

"replay function" of an initial sensory trace, imagery supposedly activates sensory representations without prior stimulation. This means that imagery solely makes use of long-term memory contents and constitutes a form of memory recollection in the perceptual domain. Studying the similarity of timbre imagery and perception, Halpern et al. (2004) had musicians rate perceived dissimilarity of subsequently presented pairs of timbres while recording brain activity with functional magnetic resonance imaging. The same procedure (including the dissimilarity ratings) was repeated in a condition in which the auditory stimuli were to be actively imagined. Figure 4.4 depicts the significant correlation between the behavioral dissimilarity data in the perception and imagery conditions. When compared to a visual imagery control condition, both auditory perception and imagery conditions featured activity in the primary and secondary auditory cortices with a right-sided asymmetry. Results such as these speak for the accuracy of auditory imagery for timbre: Sensory representations activated by imagery can resemble those activated by sensory stimulation.

These empirical findings have a bearing on the conceptualization of the active facets of timbre cognition. Working memory for timbre seems to be characterized as relying on concrete sensory refreshing or re-enactment and differs from the motor-based articulation processes found for pitch and verbal memory. Auditory imagery based on LTM representations of timbre appears to accurately resemble actual sensory stimulation. Both processes, refreshing and imagery, are related to the notion of *active perceptual simulation*, which is defined as a re-creation of facets of perceptual experience. Theories of perceptual symbol systems advocate that cognition is grounded in perceptual simulation (Barsalou 1999). This view stands in direct contrast to classic theories of cognition, which presume that perceptual processing leads to a transduction of sensory states into configurations of amodal symbols (Atkinson and Shiffrin 1968). Perceptual symbol systems assume that sensory schemata are abstracted from sensory states via perceptual learning, and cognition consists of simulating these schematic representations in concrete sensory form. That framework would be able to account for this phenomenon: When listeners actively maintain timbre in WM, they "hear" the original sound. Similarly, when a conductor reads a score, they will not perceive the music through the abstract application of a set of music-theoretical rules but through the mental restaging of the notated musical scene (cf., Zatorre and Halpern 2005).

4.5 Interference Effects in Memory for Timbre

An important part of the characterization of auditory memory concerns the question of whether timbre is encoded and stored independently from other auditory attributes. In this section, three specific scenarios will be described that address aspects of interference in short-term memory for timbre, effects of musical timbre on long-term melodic memory, and effects of voice timbre on verbal memory.

4.5.1 Interference in Short-Term Memory

Short-term storage of timbre is closely related to the perceptual encoding stage. In basic perceptual experiments that have tested the independence of pitch and timbre, results indicate that pitch and timbral brightness information are integral attributes (Melara and Marks 1990; Allen and Oxenham 2014; and for a more detailed discussion, see McAdams, Chap. 2). There is evidence to suggest that interactions between pitch and timbre extend to memory.

Siedenburg and McAdams (2018) studied the short-term recognition of timbre by using a serial matching task wherein participants judged whether the timbres of two subsequent (standard and comparison) sequences of tones were of the same order. When the tone sequences comprised concurrent variation in pitch, the performance of nonmusicians was impaired more strongly than was the performance of musicians. When pitch patterns differed across standard and comparison sequences, however, musicians showed impaired performances as well. This means that musicians may require higher degrees of complexity of pitch patterns in order to exhibit impaired timbre recognition. More generally speaking, these results indicate that pitch and timbre are not encoded independently in short-term memory—these features are part of an integrated memory trace.

The topic of pitch-timbre interference implies an answer to the question of whether the defining units of working memory are constituted by integrated auditory events (called *sound objects*) or by individual features. Joseph et al. (2015) investigated the recognition of narrowband noise segments. Two features of these

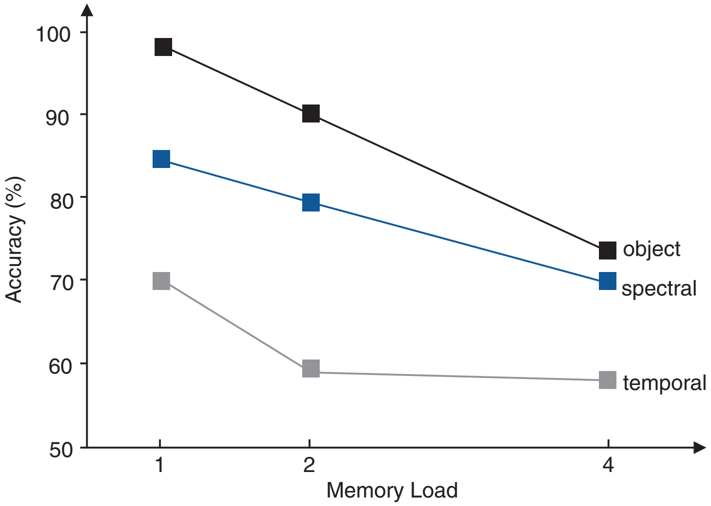

Fig. 4.5 Accuracy by memory load and condition in the item recognition task. Participants were required to match a probe sound to sounds from a previously presented sequence of length 1, 2, or 4 (*Memory Load*) according to *spectral*, *temporal*, or both spectral and temporal features (*object*). (Recreated from Joseph et al. 2015; used with permission from Frontiers/Creative Commons)

sounds were manipulated in the experiment: the spectral passband (i.e., yielding differences in the spectral centroid) and the amplitude modulation (AM) rate imposed on the waveform. Listeners were presented with a sequence of three sounds (each of 1 s duration with 1 s inter-stimulus intervals). They were instructed to judge whether there was a match between the third probe sound and the first or second sound presented. In two feature conditions, a match was defined as having one identical feature: passband or AM rate. In the object condition, a match was defined as both features being identical. As depicted in Fig. 4.5, accuracy in the object condition exceeded that in the feature condition by far (although accuracy for the spectral feature alone was better compared to the AM feature alone). This means that even if the task required participants only to memorize individual component features, there was a significant extraction cost when features had to be encoded and recollected individually.

Whether concerning the interference of pitch and timbre (Siedenburg and McAdams 2018) or spectral and temporal features of noise realizations (Joseph et al. 2015), the empirical evidence indicates that the content of short-term storage appears to be integrated auditory events (or "objects" as termed by Joseph et al. 2015) rather than individual features. The same position will be corroborated in the following review of effects of timbre on memory for melodies.

4.5.2 Timbre and Long-Term Melodic Memory

This section summarizes studies that have investigated the effects of timbre on melodic memory at time spans in the range of at least several minutes, which generally would be considered as LTM rather than STM processes. Although timbre does not affect a melody's pitch and rhythm structure, many studies have highlighted the role of timbre as a salient auditory feature for memorizing melodies. In experiments by Radvansky et al. (1995), participants identified which of two test melodies, a target and a distractor, was heard in the experiment's exposure phase. The accuracy of recognition judgements by both musicians and nonmusicians was higher when the timbre of the test melody equaled the timbre of the exposure melody, that is, a change in instrumentation clearly impaired melody recognition. This result was replicated with a sample of 6-month-old infants (Trainor et al. 2004).

Using richer musical stimuli, Poulin-Charronnat et al. (2004) studied recognition memory for tonal music (Liszt) and atonal contemporary music (Reynolds). A change of instrumentation from piano to orchestra or vice versa impaired recognition of tonal excerpts in both musicians and nonmusicians compared to conditions in which the instrumentation was held constant. For contemporary music, recognition performance by musicians was strongly impaired for instrumentation changes, whereas there was no effect for nonmusicians who performed poorly regardless of instrumentation. Halpern and Müllensiefen (2008) observed that the detrimental effect of timbre change is unaffected by whether the participant's attention at the exposure stage was directed toward timbral features (through an instrument categorization task) or to the melodic structure (through a judgement of melody familiarity).

Most recently, Schellenberg and Habashi (2015) explored the temporal dynamics of musical memory by testing melody recognition with delays between the exposure and the test that spanned 10 min, 1 day, and 1 week. Surprisingly, recognition accuracies were similar for all three retention intervals, and there even seemed to be a trend for consolidation as reflected by a small but significant increase in accuracy for a delay of 1 week compared to 10 min. Pitch transpositions of six semitones or a tempo shift of sixty-four beats per minute impaired recognition after 10 min and 1 day but not after 1 week. Notably, a change of instrument from piano to saxophone impaired melody recognition as strongly as the aforementioned changes in pitch or tempo but, unlike these parameters, the effect of timbre change did not reduce over time. This means that in contrast to key or tempo shifts, timbre information was not abstracted over time but stayed integral to the identity of the melody.

Schutz et al. (2017) considered melodic memory and object-to-melody association with a specific focus on the role of the amplitude envelopes of tones, which are closely related to the ways in which a sounding object is set into vibration. The excitations of a resonator by an impact usually generate rapid increases and exponentially decaying amplitude envelopes, whereas continuous excitations generate amplitude envelopes that tend to be rather flat. Schutz et al. (2017) let participants listen to melodies consisting of four pure tones with a flat or an exponentially decaying envelope. Each melody was presented three times and listeners were asked to associate the melody with a household object (e.g., digital clock, keys, calculator, etc.) that was physically presented by the experimenter during the presentation of the melodies. After a delay of more than 6 min, participants were presented with a recognition and recollection task; if melodies were identified as old, listeners also were asked to recall the associated object. Although their results only exhibited insignificant trends toward better melody recognition for percussive envelopes, melody-to-object association was significantly better for tones with percussively decaying envelopes. In two additional experiments, the authors observed that melodies of tones with reverse-ramped (i.e., increasing) envelopes were poorly associated with objects (performance was even worse than with flat envelopes). The results indicated that associative memory was better for decaying envelopes compared to flat or reversed envelopes, potentially due to their higher ecological familiarity. Although it may not be clear a priori why this stimulus manipulation only had an effect on associative memory but not on recognition memory, differences between associative and recognition memory are frequently observed in the literature (Kahana 2012).

Taken together, these studies strongly suggest that memory for melodies does not solely draw from an abstract lexicon of melodies represented by pitch interval information. Instead, melody recognition appears to rely on a rich auditory representation that integrates various features including timbre. Similar results have been found for verbal memory as described in the next section.

4.5.3 Timbre and Verbal Memory

The classic study on the role of voice timbre in spoken word recognition was conducted by Goldinger (1996) (for a discussion of more recent studies, see Goh 2005). Goldinger (1996) let participants listen to sequences of words recorded by 2, 6, or 10 different speakers. After three different delay periods, participants were required to distinguish old from new words in a recognition task. The results indicated that listeners better recognized words spoken by the voices of the exposure phase: the same-voice advantage was 7.5% after 5 min, 4.1% after 1 day, and an unreliable 1.6% after 1 week. Beyond the coarse same/different distinction, however, there was also a more fine-grained correlation of voice similarity with the percentage of correct rejections. In a second experiment, the delay interval was held constant at 5 min, but there were three different encoding conditions. Using a speeded classification task, participants either categorized voice gender, the initial phoneme from a list of alternatives, or the word's syntactic category (e.g., verb versus adjective). From a levels-of-processing perspective (Craik and Lockhart 1972), these tasks enforce shallow (gender), intermediate (phoneme), or deep (syntax) encoding of the words, respectively. The word recognition scores were as expected in that hit rates increased with the depth of encoding (i.e., gender < phoneme < syntax). The strength of the voice effect was reversed across encoding conditions. Whereas old voices had an advantage of around 12% for the gender condition, this advantage shrank to around 5% in the syntax condition. Because the effects were robust to a variety of encoding conditions, Goldinger (1996) concluded that the results "support an episodic view of the lexicon, in which words are recognized against a background of countless, detailed traces. Speech is not a noisy vehicle of linguistic content; the medium may be an integral dimension of later representation" (p. 1180). These findings suggest that the long-standing idea of the *mental lexicon* (Oldfield 1966), supposedly based on an amodal representation of words, is not enough to account for human recognition of spoken words.

Van Berkum et al. (2008) specifically investigated the time course of the integration of speaker and message information. In their experiment, participants passively listened to sentences while electroencephalography (EEG) signals were recorded. In two anomalous conditions, sentences could either feature *semantic anomalies* (e.g., Dutch trains are *sour* and blue; target word in italic) or *speaker inconsistencies* (e.g., I have a large *tattoo* on my back, spoken with an upper-class accent). They found that semantic anomalies elicited a standard N400 response for deviant trials, that is, an inflection of the EEG signal with a negative peak around 400 ms after the target word. Interestingly, the same time course was observed for the speaker inconsistency condition, where a similar N400 response was observed (albeit of much smaller magnitude). The clear onset of the deviant EEG response at around 200–300 ms after the acoustic onset of the deviant word indicated the rapid extraction and processing of timbre-specific information. These results suggest that voice-specific information is integrated into linguistic processing around the same point in time when language interpretation mechanisms construct meaning based on the lexical content of the words.

In sum, the studies presented in this section have shown strong associations between individual features in auditory memory. Because of their shared underlying tonotopic dimension, pitch and timbral brightness may be particularly intertwined. However, some of the evidence suggests that even amplitude envelope features affect aspects of melodic memory (Schutz et al. 2017). If features are to be accessed, recollected, or recognized individually, an extraction cost can be assumed (Joseph et al. 2015). This cost may be reduced by enhanced auditory attention and listening experience to some extent, but it is unlikely to ever vanish completely (Allen and Oxenham 2014; Siedenburg and McAdams 2018). The notion of integrated memory representations appears to contradict the seemingly abstract nature of auditory cognition (e.g., Obleser and Eisner 2009; Patel 2008). Sensory information related to timbre is not simply "left behind" in the process of information transduction from sensory to more symbolic forms of representations. On the contrary, timbre stays integral to both word and melody recognition over long retention spans—the medium and the message are intertwined.

4.6 Familiarity and Voice Superiority

The last theme in this review of memory for timbre concerns the roles of long-term familiarity with sound sources. A sound source of particular relevance and familiarity for humans is the voice. For that reason, the role of the voice in timbre processing has been studied with particular scrutiny (for an overview, see Mathias and Kriegstein, Chap. 7). This section discusses studies that have investigated the role of long-term familiarity with musical-instrument sounds in timbre processing (Sect. 4.6.1) and the special status of voice timbre in melodic memory (Sect. 4.6.2).

4.6.1 Familiarity in Short-Term Recognition

A factor that significantly increases the complexity of STM research and modeling relates to the presumption that STM is not completely distinct from LTM as suggested by procedural memory approaches. In fact, there is further evidence to assume a strong link between the two systems (e.g., Jonides et al. 2008). The experimental cornerstone regarding this link in verbal memory research is the *lexicality effect*: short-term memory for the identity of words or syllables (i.e., verbal items) is generally better for words than for *pseudowords* or nonsense syllables (Thorn et al. 2008). Pseudowords are defined as meaningless strings of letters that respect a language's phonotactic constraints but are not part of the dictionary (e.g., bech, chaf, tog, wesh, etc.).

Similar enhancements of STM performance have also been demonstrated for related linguistic variables, including word frequency and imaginability (Thorn et al. 2008). The analogous question for timbre, and of particular concern for the current purpose, is whether STM is facilitated by long-term familiarity with sounds produced by well-known musical instruments. If this were the case, it would constitute a timbral analogy to the verbal lexicality effect. More importantly, it would suggest that STM for timbre cannot be properly placed in a one-size-fits-all principle of sensory persistence—one would need to consider existing auditory categories as well.

To study the role of familiarity in STM for timbre, Siedenburg and McAdams (2017) compared the recognition of recorded tones from familiar acoustical instruments with that of unfamiliar synthesized tones that do not readily evoke sound-source categories. Steps were taken in order to manipulate familiarity while controlling for dissimilarity relations within the stimulus set. First, the spectrotemporal signal envelopes and temporal fine structures of recorded sounds were mismatched to generate novel and unfamiliar sounds. Second, familiarity ratings by musicians were collected for the transformed sounds, ensuring that the transformed sounds used in the main experiment were rated as significantly less familiar compared to the original recordings. Third, the main experiment used an item recognition task with sequences of three sounds. The mean timbral dissimilarity between the sounds in the sequence and those in the probe was equalized across recordings and transformations, using previously obtained pairwise dissimilarity ratings. Two experiments revealed greater recognition accuracy for timbres of familiar recorded sounds compared to unfamiliar transformations, as well as better performance at shorter delays (2 s versus 6 s), but no interaction between the factors of delay and stimulus material. These results point toward a generally more robust form of encoding of timbral properties coming from familiar acoustical instruments. The superior memory performance for familiar instruments proved to be independent of effects of perceptual similarity.

Prior knowledge of instrument categories for familiar acoustical-instrument sounds helps to associate sounds with auditory knowledge categories or schemas. In other words, familiar instrument sounds activate not only auditory sensory representations but, possibly to some extent, also activate semantic, visual, and even sensorimotor networks. These sounds are not necessarily rehearsed in STM, but could act as representational anchors for the associated auditory sensory traces. Saitis and Weinzierl (Chap. 5) further describe the nuanced cross-modal associations that timbre can elicit.

The special role of sound source familiarity has gained support from neurophysiological studies on timbre processing. Pantev et al. (2001) observed that professional trumpet players and violinists exhibited stronger event-related potentials to sounds from their own instrument at around 100 ms after sound onset (the N1 component), indexing stronger pre-attentive processes related to stimulus detection. In addition, there is evidence that learning not only affects cortical activity but can even modulate low-level processing in the brainstem. Strait et al. (2012) demonstrated that recordings of electrical brainstem activity taken from pianists more closely correlated with the amplitude envelopes of the original piano sounds when compared to recordings taken from musicians who did not play the piano as their primary instrument. However, brainstem activity did not differ between pianists and other musicians for sounds from the tuba and the bassoon. This result indicates that there may be instrument-specific neural adaptations that affect the perceptual processing of certain classes of instrumental sounds. Apparently, musical training can affect the fine-tuning of subcortical structures to more efficiently process sounds that are of particular relevance to the listener. These findings refute the idea that timbre could be a less important auditory surface feature. On the contrary, elementary aspects of auditory processing appear to be shaped by experience with sound source categories.

Unfortunately, none of the studies discussed here have been able to completely control low-level factors and the individual experience of the participants. Therefore, the exact origins of the effects may remain contentious. Future experiments that familiarize listeners with certain classes of novel timbres in the lab may help to more precisely characterize the underlying mechanisms of familiarity in timbre processing.

4.6.2 Voice Superiority

A sound source that all humans should be particularly familiar with, from both an evolutionary and ontogenetic point of view, is the human voice. Recent studies have suggested that sounds of vocal origin are faster and more robustly categorized compared to instrumental musical sounds. Many of these studies are also discussed in greater depth by Agus, Suied, and Pressnitzer (Chap. 3); hence, they will only be summarized here to set the stage for the consideration of additional memory effects.

Employing a go/no-go task, Agus et al. (2012) asked listeners to indicate as quickly as possible whether sounds were part of a target category (voice, percussion, or strings). Results showed faster reaction times for voices. Importantly, the effect did not arise for auditory chimeras that retained either spectral or temporal envelope shapes of vocal sounds. Suied et al. (2014) further observed that voices were more robustly recognized compared to other instrumental sounds even for very short snippets (durations from 2 ms to 128 ms). The exact acoustic features responsible for this advantage must be of spectrotemporal nature because neither solely spectral nor solely temporal cues sufficed to yield a processing advantage. Furthermore, Agus et al. (2017) only observed an increase of activity in areas of the human temporal lobe that have documented sensitivity to vocal stimuli (see Mathias and Kriegstein, Chap. 7) for nonchimaeric stimuli. This means that there are brain areas that selectively react to the full set of spectrotemporal cues of voices but not to isolated spectral or temporal cues.

Across several recent studies, Weiss and colleagues (see Weiss et al. 2017, and references therein) accumulated evidence for a memory advantage of vocal melodies compared to melodies played by nonvocal musical instruments (specifically piano, banjo, and marimba). In all of these studies, the basic experimental approach

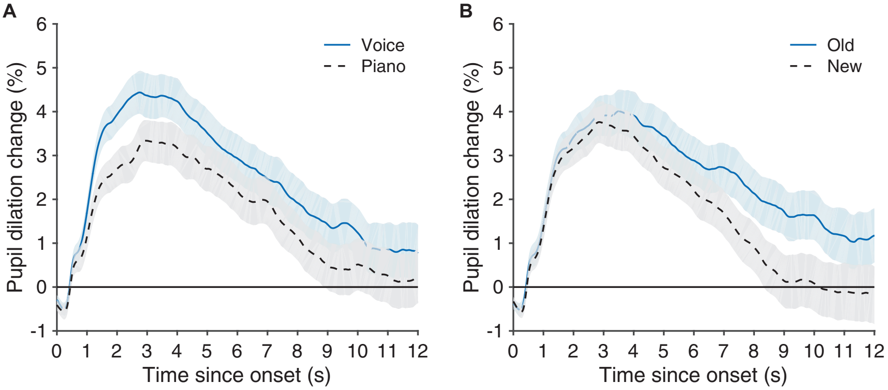

Fig. 4.6 Pupil dilation response as a function of the time since melody onset: (**A**) vocal versus piano melodies; (**B**) old versus new melodies. (From Weiss et al. 2016; used with permission of the American Psychological Association)

was to have participants listen to a set of melodies presented with a vocal or instrumental timbre. After a 5–10 min break, participants heard the exposure melodies intermixed with a set of new melodies and rated their confidence in having heard a melody previously on a seven-point scale. Analyses of the recognition ratings for old and new melodies revealed that adults more confidently and correctly recognized vocal compared to instrumental melodies (Weiss et al. 2012). The effect generalized to musicians with and without absolute pitch, and even pianists recognized more vocal melodies correctly with higher confidence in their correct ratings than for piano melodies (Weiss et al. 2015). This finding suggests that sensorimotor representations and perceptual familiarity with certain classes of sounds are an unlikely locus of the observed effect. Otherwise, pianists should have shown a reduced voice advantage due to their ability to recruit motor representations for piano melodies and to their high familiarity with piano sounds.

It was further shown that the presentation of vocal melodies, as well as previously encountered melodies, was accompanied by an increase in pupil dilation (Weiss et al., 2016). Increases in pupil dilation are generally interpreted as an indicator of heightened engagement and potentially a greater recruitment of attentional resources (Kang et al. 2014). The results by Weiss et al. (2016) are depicted in Fig. 4.6. Note that the difference in pupil dilation between piano and vocal melodies is most pronounced around 3 s after the onset of melodies. To the contrary, the difference between old and new melodies appears to accumulate across the full length of the melodies, indexing the distinct time courses of melody recognition and vocal superiority.

Although the memory advantage for melodies with a vocal timbre has turned out to be stable across several studies, there remain several open questions to explore within this paradigm (e.g., the role of signal amplitude normalizations, see Bigand et al. 2011). Most importantly, the psychophysical origin of any of the reported vocal superiority effects (Agus et al. 2012; Weiss et al. 2012) is not clear. Could vocal superiority be a result of the involvement of motor processes (Liberman and

Mattingly 1985)? Is there a particular spectrotemporal feature in the acoustics of voices that boosts the processing of these sounds? Or is it the case that all auditory stimuli that indicate a vocal sound source happen to be preferentially processed once a voice has been implicitly recognized? Differentiating these hypotheses would require disentangling top-down and bottom-up effects. As discussed in greater depth by Mathias and Kriegstein (Chap. 7), there are voice-selective areas in the auditory cortex that only react to vocal input sounds, even if low-level cues, such as temporal or spectral envelopes, are matched with other sounds (Agus et al. 2017). But what exactly is the representational content of these voice-selective areas? Is this cortical selectivity the origin or the result of vocal superiority? Future research may be able to shed light on these intriguing questions.

4.7 Summary and Future Perspectives

This chapter provides a review of important research threads in memory for timbre. These threads concern the role of perceptual similarity relations and chunking in short-term memory for timbre, active imagery of timbre, the role of interference of auditory attributes in memory, and questions regarding the privileged processing of familiar and vocal timbres. Only 10 years ago these topics had not been covered to any serious degree within auditory cognition research. Since then, many studies have been published that provide valuable insights into the processing of timbre in memory, but they also open up new perspectives for future research. Today, we think we have sufficient empirical grounds to formulate a few principles of how memory for timbre works. In the following, five such principles will be outlined, followed by a brief discussion of what we consider to be relevant questions for future research.

4.7.1 Principles of Memory for Timbre

In contrast to other sets of memory principles that have been proposed to hold for all types memory (Surprenant and Neath 2009), the current principles are specifically derived from empirical studies on timbre and they serve two purposes. First, these principles will act as concise summaries of the empirical data collected up to date. Second, they will be considered as intermediate explanations of empirical effects. From this perspective, a principle should be more abstract than an effect. At the same time, a principle can be less specific than a model because it does not need to provide a comprehensive list of components and their functional interrelations for the overall system. In this sense, the following principles highlight what we currently understand about memory for timbre but also expose how incomplete the current state of knowledge is. Figure 4.7 provides a schematic of how these processes could function for the example of an item recognition task.

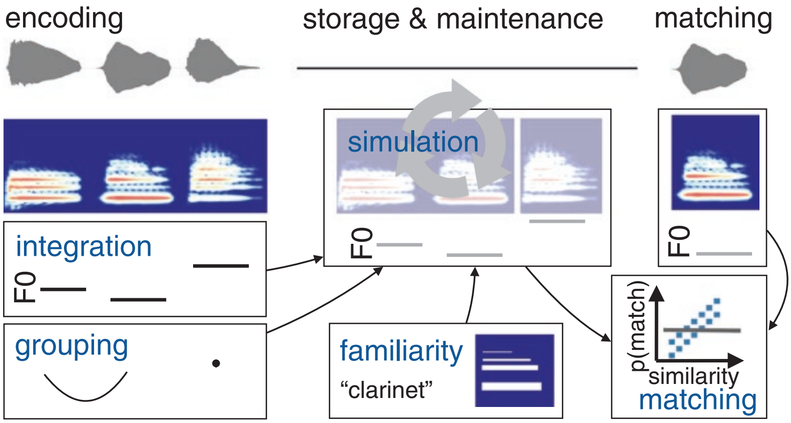

Fig. 4.7 Schematic of the five proposed principles of memory for timbre. The example shows how the five principles might relate to each other in a timbre memory recognition task. The auditory spectrogram indicates that the timbres of a sequence of sounds are represented in terms of their spectrotemporal properties. The structure of the memory trace is shaped by the process of *integration* (Principle I) as concurrently varying features, such as pitch, are integrated with the timbre memory trace. Sequential *grouping* (Principle II) provides additional temporal structure to the memory trace (in this example by separating the last sound from the first two). Timbre *familiarity* (Principle III) provides representational anchor points and cross-modal associations, for instance, by readily yielding semantic labels for certain sounds (here, the clarinet). Attention-based refreshing, a form of perceptual *simulation* (Principle IV), may be a maintenance strategy specifically suited for timbre. Here, perceptual simulation is graphically represented by a circle, denoting the cyclical process of refreshing the memory trace by means of attention. Finally, the *matching* (Principle V) stage takes the collection of features of a probe sound and compares them to stored memory traces. If the similarity measure exceeds the listener's internal threshold, the probe is considered a match

4.7.1.1 Integration: Timbre Information as Integrated Representations in Memory

Several experiments have shown that the perceptual discrimination of pitch and timbre (and more specifically, timbral brightness) is subject to symmetric interference effects (e.g., Allen and Oxenham 2014). As reviewed in Sect. 4.5.1, recent experiments on short-term recognition found detrimental effects of concurrent variations of irrelevant features (Joseph et al. 2015; Siedenburg and McAdams 2018) and hence suggested that integrated representations (or events/auditory objects) are stored in STM. The elaborations in Sect. 4.5.2 have illustrated that experiments on long-term melodic memory corroborated these findings. Whenever there is a shift of timbre, it is harder to discriminate new from old melodies (Schellenberg and Habashi 2015). The analogous effect even constitutes a classic effect in verbal memory: Spoken words are harder to recognize whenever they stem from a different speaker in the test phase (Goldinger 1996), and the time courses of semantic and speaker information processing are very similar (Van Berkum et al. 2008).

4.7.1.2 Grouping: Memory for Timbre Sequences is Affected by Grouping Cues

The item-to-item structure of auditory sequences strongly affects their mnemonic affordances. As reviewed in Sect. 4.3.2, hierarchically structured sequences are easier to chunk and encode compared to random sequences (Siedenburg et al. 2016). Furthermore, acoustic cues such as strong acoustic dissimilarity between statistically distinct groups of sounds may enhance the separated encoding of such groups (Tillmann and McAdams 2004). Whether based on chunking or acoustic dissimilarity, grouping cues powerfully enrich memory traces by structuring them along a hierarchy of time scales.

4.7.1.3 Familiarity: Better Memory Performance and Processing Accuracy

As discussed in Sect. 4.6, familiar sounds from well-known musical instruments are easier to recognize compared to unfamiliar transformed sounds (Siedenburg and McAdams 2017). Familiar musical-instrument sounds not only activate auditory sensory representations but to some extent also elicit semantic, visual, and even sensorimotor representations, which may act as anchors for the associated auditory sensory traces. Human voices may be considered as sound sources that are particularly familiar both ontogenetically and evolutionarily, and corresponding vocal advantage effects have been demonstrated (Agus et al. 2012; Weiss et al. 2012).

4.7.1.4 Perceptual Simulation: Active Memory Rehearsal and Timbre Imagery

As described in Sect. 4.4, short-term recognition of timbre can be impaired by attention-demanding tasks such as visual change detection (Siedenburg and McAdams 2017) or auditory imagery (Soemer and Saito 2015). Furthermore, precise timbre representations can be obtained through auditory imagery (Halpern et al. 2004), that is, through a simulation of sensory schemata from long-term memory. This means that timbre is part of an active form of auditory cognition that operates at the level of sensory representations.

4.7.1.5 Matching: Timbre Recognition via Similarity-Based Matching

The similarity effects observed by Siedenburg and McAdams (2018), as discussed in Sect. 4.3.1, suggest that a similarity-based matching mechanism could be at the basis of timbre recognition. This mechanism could be conceived as an ongoing computation of similarity of the current auditory input with past representations that are stored in memory. For item recognition tasks, the matching process could effectively be modeled as a similarity computation (Kahana 2012), indicating a match if the summed perceptual similarities of the probe item to the items in the memory sequence exceeds a certain threshold. Serial recognition tasks could be based on a matching process that computes item-wise dissimilarities between two sequences and hence corresponds to the dissimilarity of the swap criterion (Siedenburg and McAdams 2018).

4.7.2 Future Perspectives

We wish to close by discussing four potentially productive avenues for future research. Obtaining a more substantiated understanding of these questions appears to be of central importance for the topic of memory for timbre itself and might even have important implications for practical applications, such as music composition and production, sonification for human-computer interactions, and speech communication technology.

A first apparent gap in the literature concerns our knowledge about the basic memory persistence of different timbre features. For example, is a set of sounds varying along temporal features (e.g., the attack time) as easily retained in memory as sounds varying along spectral timbre features (e.g., brightness)? So far, most research has either considered minute details of spectral composition (e.g., McKeown and Wellsted 2009) or has not touched at all on the question of individual perceptual features, even if global similarity relations were considered (Golubock and Janata 2013; Siedenburg and McAdams 2017). An exception might be the experiments by Schutz et al. (2017), which indicated that flat amplitude envelopes are less well-suited for soundobject associations compared to percussive (i.e., exponentially decaying) envelopes.

Closely related to this question, and even more specific than the last point, is the need to specify the origin of vocal superiority effects. Two studies have already addressed this aspect in detail (Agus et al. 2012; Suied et al. 2014) but were not able to identify acoustic features that are specific to the vocal superiority effect. It is also not clear whether the recognition advantage observed by Weiss et al. (2012) has an acoustic or a cognitive origin. In other words, we still do not know what the basic acoustic or cognitive ingredients are that make memory for voices special.

Second, despite a plethora of memory models in other domains (e.g., Kahana 2012), there is no formal model of memory for timbre that predicts listeners' responses in memory tasks on the basis of the presented audio signals. The existence of such a model would mean a significant contribution, because it would help to make explicit the set of underlying assumptions of this research field. Perhaps the greatest hurdle for constructing a timbre memory model is the difficulty of agreeing on a signal-based representation for approximating the timbre features that are most relevant perceptually. Nonetheless, significant progress has been achieved over recent years regarding the latter (see McAdams, Chap. 2; Caetano, Saitis, and Siedenburg, Chap. 11; and Elhilali, Chap. 12).

Third, interindividual differences in memory for timbre and the role of formal musical training, as well as informal music learning, in memory for timbre have not been fully addressed yet. Whereas in timbre dissimilarity perception, musical training does not appear to affect perceptual space (McAdams et al. 1995), recognition memory for timbre may be more accurate in musicians compared to nonmusicians (Siedenburg and McAdams 2017). However, no rigorous attempt has been undertaken so far to control other individual differences that might act as confounding factors (e.g., verbal working memory, general cognitive ability). In addition, it is unclear whether the differences due to musical training observed for musicians versus nonmusicians also extend to varying levels of musical training found in the general population. It is unclear (a) how large individual differences in memory for timbre are, and (b) to what other cognitive abilities are these potential differences related. Insights regarding the latter questions might give an indication of the origin of individual differences in timbre memory. The development of a standardized test of timbre memory would represent a significant step forward in this respect. Such a test could build on existing experimental paradigms (Golubock and Janata 2013) for which factors that contribute to task difficulty have been studied already.

Finally, Agus et al. (2010) demonstrated a rapid and detailed form of implicit auditory memory for noise clips, and similar processes might be at play for the timbres of unfamiliar sound sources. Nonetheless, no study has yet addressed the time course of familiarization (i.e., learning trajectory) with sound sources. Lately, Siedenburg (2018) showed that the perception of brightness can be affected strongly by context effects. This implies that listeners not only memorize timbral associations within sequences of sounds, but the percept of a sound itself can be altered by the timbral properties of the auditory context. Hence, there exists an implicit form of memory for auditory properties, including timbre, that subconsciously affects present perceptual processing and that is in urgent need of further scientific exploration.

Acknowledgements KS was supported by a Carl von Ossietzky Young Researcher's Fellowship from the University of Oldenburg.

Compliance with Ethics Requirements Kai Siedenburg declares that he has no conflict of interest.
Daniel Müllensiefen declares that he has no conflict of interest.

References

Agus TR, Thorpe SJ, Pressnitzer D (2010) Rapid formation of robust auditory memories: insights from noise. Neuron 66:610–618

Agus TR, Suied C, Thorpe SJ, Pressnitzer D (2012) Fast recognition of musical sounds based on timbre. J Acou Soc Am 131(5):4124–4133

Agus TR, Paquette S, Suied C et al (2017) Voice selectivity in the temporal voice area despite matched low-level acoustic cues. Sci Rep 7(1):11526

Allen EJ, Oxenham AJ (2014) Symmetric interactions and interference between pitch and timbre. J Acous Soc Am 135(3):1371–1379

Andrillon T, Pressnitzer D, Léger D, Kouider S (2017) Formation and suppression of acoustic memories during human sleep. Nat Commun 8(1):179

Atkinson RC, Shiffrin RM (1968) Human memory: a proposed system and its control processes. In: Spence KW, Spence JT (eds) The psychology of learning and motivation: advances in research and theory (vol 2). Academic Press, New York, pp 89–195

Baddeley AD (2012) Working memory: theories models and controversies. Ann Rev Psy 63:1–29

Barsalou LW (1999) Perceptual symbol systems. Beh Brain Sci 22:577–660

Berz WL (1995) Working memory in music: a theoretical model. Music Percept 12(3):353–364

Bigand E, Delbé C, Gérard Y, Tillmann B (2011) Categorization of extremely brief auditory stimuli: domain-specific or domain-general processes? PLoS One 6(10):e27024

Camos V, Lagner P, Barrouillet P (2009) Two maintenance mechanisms of verbal information in working memory. J Mem Lang 61(3):457–469

Cowan N (1984) On short and long auditory stores. Psy Bull 96(2):341–370

Cowan N (2001) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Beh Brain Sci 24(1):87–114

Cowan N (2008) What are the differences between long-term short-term and working memory? Prog Brain Res 169:323–338

Cowan N (2015) Sensational memorability: working memory for things we see hear feel or somehow sense. In: Jolicoeur P, Levebre C, Martinez-Trujillo J (eds) Mechanisms of sensory working memory/ Attention and perfomance XXV. Academic Press, London, pp 5–22

Craik FI, Lockhart RS (1972) Levels of processing: a framework for memory research. J Verb Learn Verb Beh 11(6):671–684

Crowder RG (1993) Auditory memory. In: McAdams S, Bigand E (eds) Thinking in sound: the cognitive psychology of human audition. Oxford University Press, Oxford, pp 113–143

Darwin CJ, Turvey MT, Crowder RG (1972) An auditory analogue of the sperling partial report procedure: evidence for brief auditory storage. Cog Psy 3(2):255–267

D'Esposito M, Postle BR (2015) The cognitive neuroscience of working memory. Ann Rev Psych 66:1–28

Demany L, Semal C (2007) The role of memory in auditory perception. In: Yost WA, Fay RR (eds) Auditory perc of sound sources. Springer, New York, pp 77–113

Demany L, Trost W, Serman M, Semal C (2008) Auditory change detection: simple sounds are not memorized better than complex sounds. Psy Sci 19(1):85–91

Demany L, Pressnitzer D, Semal C (2009) Tuning properties of the auditory frequency-shift detectors. J Acou Soc Am 126(3):1342–1348

Demany L, Semal C, Cazalets J-R, Pressnitzer D (2010) Fundamental differences in change detection between vision and audition. Exp Brain Res 203(2):261–270

Deutsch D (1970) Tones and numbers: specificity of interference in immediate memory. Sci 168(3939):1604–1605

Dudai Y (2007) Memory: it's all about representations. In: Roediger HL III, Dudai Y, Fitzpatrick SM (eds) Science of memory: concepts. Oxford University Press, Oxford, pp 13–16

Goh WD (2005) Talker variability and recognition memory: instance-specific and voice-specific effects. J Exp Psy:LMC 31(1):40–53

Goldinger SD (1996) Words and voices: episodic traces in spoken word identification and recognition memory. J Exp Psy: LMC 22(5):1166–1183

Golubock JL, Janata P (2013) Keeping timbre in mind: working memory for complex sounds that can't be verbalized. J Exp Psy: HPP 39(2):399–412

Hagoort P, Indefrey P (2014) The neurobiology of language beyond single words. Ann Rev Neuosci 37:347–362

Halpern AR, Müllensiefen D (2008) Effects of timbre and tempo change on memory for music. Q J Exp Psy 61(9):1371–1384

Halpern AR, Zatorre RJ, Bouffard M, Johnson JA (2004) Behavioral and neural correlates of perceived and imagined musical timbre. Neuropsy 42(9):1281–1292

James W (1890/2004) The principles of psychology. http://www.psychclassicsyorkuca/James/Principles. Accessed 9 Nov 2015

Jonides J, Lewis RL, Nee DE et al (2008) The mind and brain of short-term memory. Ann Rev Psy 59:193–224

Joseph S, Kumar S, Husain M, Griffiths T (2015) Auditory working memory for objects vs features. Front Neurosci 9(13). https://doi.org/10.3389/fnins201500013

Kaernbach C (2004) The memory of noise. Exp Psy 51(4):240–248

Kahana MJ (2012) Foundations of human memory. Oxford University Press, New York

Kang OE, Huffer KE, Wheatley TP (2014) Pupil dilation dynamics track attention to high-level information. PLoS One 9(8):e102463

Lerdahl F, Jackendoff R (1983) A generative theory of tonal music. MIT Pr, Cambridge

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21:1–36

Ma WJ, Husain M, Bays PM (2014) Changing concepts of working memory. Nat Neurosci 17(3):347–356

McAdams S, Winsberg S, Donnadieu S et al (1995) Perceptual scaling of synthesized musical timbres: common dimensions specificities and latent subject classes. Psy Res 58(3):177–192

McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. Nat Neurosci 16(4):493–498

McKeown D, Wellsted D (2009) Auditory memory for timbre. J Exp Psy: HPP 35(3):855–875

McKeown D, Mills R, Mercer T (2011) Comparisons of complex sounds across extended retention intervals survives reading aloud. Perception 40(10):1193–1205

Melara RD, Marks LE (1990) Interaction among auditory dimensions: timbre pitch and loudness. Perc Psyphys 48(2):169–178

Meyer LB (1956) Emotion and meaning in music. Chicago U Pr, Chicago

Miller GA (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. Psy Rev 63(2):81–97

Müllensiefen D, Halpern AR (2014) The role of features and context in recognition of novel melodies. Music Percept 31(5):418–435

Nees MA, Corrini E, Leong P, Harris J (2017) Maintenance of memory for melodies: articulation or attentional refreshing? Psy Bull Rev 24(6):1964–1970

Obleser J, Eisner F (2009) Pre-lexical abstraction of speech in the auditory cortex. Tr Cog Sci 13(1):14–19

Oldfield RC (1966) Things words and the brain. Q J Exp Psy 18(4):340–353

Pantev C, Roberts LE, Schulz M et al (2001) Timbre-specific enhancement of auditory cortical representations in musicians. Neur Rep 12(1):169–174

Patel AD (2008) Music language and the brain. Oxford University Press, Oxford

Poulin-Charronnat B, Bigand E, Lalitte P et al (2004) Effects of a change in instrumentation on the recognition of musical materials. Music Percept 22(2):239–263

Radvansky GA, Fleming KJ, Simmons JA (1995) Timbre reliance in non-musicians' and musicians' memory for melodies. Music Percept 13(2):127–140

Saffran JR, Johnson EK, Aslin RN, Newport EL (1999) Statistical learning of tone sequences by human infants and adults. Cogn 70:27–52

Saxena SK (2008) The art of Tabla rhythm: essentials tradition and creativity. In: New vistas in Indian performing arts. DK Printworld Ltd, New Dehli

Schellenberg EG, Habashi P (2015) Remembering the melody and timbre forgetting the key and tempo. Mem Cog 43(7):1021–1031

Schulze K, Koelsch S (2012) Working memory for speech and music. A NY Ac Sci 1252(1):229–236

Schulze K, Tillmann B (2013) Working memory for pitch timbre and words. Mem 21(3):377–395

Schutz M, Stefanucci JK, Baum SH, Roth A (2017) Name that percussive tune: Associative memory and amplitude envelope. Q J Exp Psy 70(7):1323–1343

Siedenburg K (2018) Timbral Shepard-illusion reveals perceptual ambiguity and context sensitivity of brightness perception. J Acou Soc Am 143(2):EL-00691

Siedenburg K, McAdams S (2017) The role of long-term familiarity and attentional maintenance in auditory short-term memory for timbre. Mem 25(4):550–564

Siedenburg K, McAdams S (2018) Short-term recognition of timbre sequences: music training pitch variability and timbral similarity. Music Percept 36(1):24–39

Siedenburg K, Mativetsky S, McAdams S (2016) Auditory and Verbal Memory in North Indian Tabla Drumming. Psychomusicology 26(4):327–336

Simon HA (1978) Information-processing theory of human problem solving. In: Estes WK (ed) Handbook of learning and cognitive processes, vol 5, pp 271–295

Soemer A, Saito S (2015) Maintenance of auditory-nonverbal information in working memory. Psy Bull Rev 22(6):1777–1783

Starr GE, Pitt MA (1997) Interference effects in short-term memory for timbre. J Acou Soc Am 102(1):486–494

Strait DL, Chan K, Ashley R, Kraus N (2012) Specialization among the specialized: auditory brainstem function is tuned in to timbre. Cortex 48(3):360–362

Suied C, Agus TR, Thorpe SJ et al (2014) Auditory gist: recognition of very short sounds from timbre cues. J Acou Soc Am 135(3):1380–1391

Surprenant A, Neath I (2009) Principles of memory. Psy Pr, New York

Thorn AS, Frankish CR, Gathercole SE (2008) The influence of long-term knowledge on short-term memory: evidence for multiple mechanisms. In: Thorn AS, Page M (eds) Interactions between short-term and long-term memory in the verbal domain. Psy Pr, New York, pp 198–219

Tillmann B, McAdams S (2004) Implicit learning of musical timbre sequences: statistical regularities confronted with acoustical (dis)similarities. J Exp Psy: LMC 30(5):1131–1142

Trainor LJ, Wu L, Tsang CD (2004) Long-term memory for music: infants remember tempo and timbre. Dev Sci 7(3):289–296

van Berkum JJ, van den Brink D, Tesink CM et al (2008) The neural integration of speaker and message. J Cog Neurosci 20(4):580–591

Visscher KM, Kaplan E, Kahana MJ, Sekuler R (2007) Auditory short-term memory behaves like visual short-term memory. PLoS Bio 5(3):e56. https://doi.org/10.1371/journal.pbio.0050056

Weiss MW, Trehub SE, Schellenberg EG (2012) Something in the way she sings enhanced memory for vocal melodies. Psy Sci 23(10):1074–1078

Weiss MW, Vanzella P, Schellenberg EG, Trehub SE (2015) Pianists exhibit enhanced memory for vocal melodies but not piano melodies. Q J Exp Psy 68(5):866–877

Weiss MW, Trehub SE, Schellenberg EG, Habashi P (2016) Pupils dilate for vocal or familiar music. J Exp Psy: HPP 42(8):1061–1065

Weiss MW, Schellenberg EG, Trehub SE (2017) Generality of the memory advantage for vocal melodies. Music Percept 34(3):313–318

Zatorre RJ, Halpern AR (2005) Mental concerts: musical imagery and auditory cortex. Neur 47(1):9–12

Chapter 5
The Semantics of Timbre

Charalampos Saitis and Stefan Weinzierl

Abstract Because humans lack a sensory vocabulary for auditory experiences, timbral qualities of sounds are often conceptualized and communicated through readily available sensory attributes from different modalities (e.g., bright, warm, sweet) but also through the use of onomatopoeic attributes (e.g., ringing, buzzing, shrill) or nonsensory attributes relating to abstract constructs (e.g., rich, complex, harsh). The analysis of the linguistic description of timbre, or timbre semantics, can be considered as one way to study its perceptual representation empirically. In the most commonly adopted approach, timbre is considered as a set of verbally defined perceptual attributes that represent the dimensions of a semantic timbre space. Previous studies have identified three salient semantic dimensions for timbre along with related acoustic properties. Comparisons with similarity-based multidimensional models confirm the strong link between perceiving timbre and talking about it. Still, the cognitive and neural mechanisms of timbre semantics remain largely unknown and underexplored, especially when one looks beyond the case of acoustic musical instruments.

Keywords Auditory roughness · Auditory semantics · Cognitive linguistics · Conceptual metaphor · Crossmodal correspondence · Describing sound · Magnitude estimation · Musical meaning · Qualia · Semantic differential · Sound color · Sound mass · Sound quality · Timbral brightness · Verbal attribute

5.1 Introduction

> After consultations with his teacher and with the great violinist and collector Efrem Zimbalist … Yehudi [Menuhin] played on all three [Stradivari violins] and opted for the "Khevenhüller." (As a test piece he played "The Prayer" from Handel's Dettingen *Te Deum*.). It was to be his principal instrument for over 20 years. He described it as "ample and round, varnished in a deep, glowing red, its grand proportions … matched by a sound

C. Saitis (✉) · S. Weinzierl
Audio Communication Group, Technische Universität Berlin, Berlin, Germany
e-mail: charalampos.saitis@campus.tu-berlin.de; stefan.weinzierl@tu-berlin.de

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_5

> at once powerful, mellow and sweet.” Antonio Stradivarius had made the instrument in 1733, his 90th year, when despite his advancing years he was still at the peak of his powers (Burton 2016, p. 86).

What is a mellow and sweet sound? Imagine yourself listening to a recording of the famous violinist Yehudi Menuhin (1916–1999) performing on his Khevenhüller Strad. How would you describe the *sound* of the violin or the *sound* of Menuhin? What about the *sound quality* of the recording? Musicians, composers, sound artists, listeners, acousticians, musical instrument makers, audio engineers, scholars of sound and music, even sonar technicians, all share a subtle vocabulary of verbal attributes when they need to describe timbral qualities of sounds. These verbalizations are not crucial for processing timbre—listeners can compare (McAdams, Chap. 2), recognize (Agus, Suied, and Pressnitzer, Chap. 3), or memorize and imagine (Siedenburg and Müllensiefen, Chap. 4) timbral qualities without having to name them (Wallmark 2014). However, the way we talk about sensory experiences can disclose significant information about the way we perceive them (Dubois 2000; Thiering 2015). Menuhin’s mellow and sweet sound is a particular *concept*, an abstract yet structured idea anchored to and allowing one to make sense of a particular perceptual representation (Wallmark 2014). As such, a relation must exist between the physical properties of a sound that give rise to timbre and its semantic description.

Results of multidimensional scaling of pairwise sound dissimilarity ratings (McAdams, Chap. 2) usually show that timbre may be adequately explained on the basis of just two or three dimensions; a number many times smaller than the plethora of words and phrases used to communicate timbral impressions. On the one hand, this might be due to specific perceptual features of individual sounds (referred to as *specificities*) that are not mapped onto the shared dimensions of the prevailing timbre space. For example, the suppression of even harmonics in clarinet tones, which typically elicits an impression of hollowness, was not accounted for by classic geometric timbre models alone (e.g., McAdams et al. 1995). On the other hand, individual verbalizations can be thought of as representing *microconcepts*—basic elements of semantic knowledge activated by a stimulus object that are not fully meaningful on their own but instead yield meaning when assembled into broader semantic categories (Saitis et al. 2017). Among the diverse timbre vocabulary, therefore, many seemingly unassociated words may share the same meaning and refer to the same perceptual dimension.

Accordingly, the main goals of the research ideas and tools discussed in this chapter are twofold: to identify the few salient semantic substrates of linguistic descriptions of timbral impressions that can yield consistent and differentiating responses to different timbres along with their acoustic correlates and to quantify the relationship between perceptual (similarity-based) and semantic (language-based) representations for timbre. Important questions include the following:

- How similar are semantic timbre spaces between different categories of sound objects, for example, between instrument families and between instruments, voices, and nonmusical sounds?

- Do timbre verbalizations rely explicitly on acoustic cues or are they subject to source-cause categorical influences?
- Are timbre verbalizations a product of cultural dependencies or is timbre semantics cross-cultural?
- What are the neurobiological mechanisms underlying timbral semantic processing?
- In what ways does timbre contribute to larger-scale musical meaning?
- What is the relation between emotion and the semantics of timbre?

Subsequent sections attempt to address these questions. Section 5.2 examines how different communities of listeners verbally negotiate sound qualities and the underlying conceptualizations of timbre. In general, verbal attributes of timbre are predominantly metaphorical in nature, and empirical findings across different types of sounds and analytical approaches converge to a few salient semantic substrates, which are not very different from early theorizations for a low-dimensional semantic space of timbre by Stumpf (1890) and Lichte (1941). These findings are described in Sect. 5.3 and examined further in Sect. 5.4 through psychophysical investigations and interlanguage comparisons.

As with most aspects of timbre, much work on timbre semantics has investigated acoustic musical instruments by means of recorded samples or synthetic emulations. However, talking about instrumental timbre always implicates the acoustic environment in which the instrument is heard. In what ways do the semantics of spaces interact with the semantics of timbre? A preliminary discussion on this important but understudied question is given in Sect. 5.5. Finally in Sect. 5.6, overarching ideas are summarized and new directions for future research are proposed.

Two considerations are necessary before proceeding. First, sound source identification (e.g., this is not a violin) is in itself a type of timbre semantics. The consistent use of onomatopoeia in verbal descriptions of musical and environmental timbres (see Sect. 5.2.1) is one example of identification acting as semantics. In practice, however, timbre semantics is typically defined as *qualia* (this chapter) and sound source perception is studied separately (see McAdams, Chap. 2; Agus, Suied, and Pressnitzer, Chap. 3). Second, in studying timbre semantics as qualia, a distinction will be made between *timbre* as sound quality of complex spectra (this chapter) and *sound quality* as an evaluation of functionality and pleasantness in audio reproduction and industrial sound design contexts (see Lemaitre and Susini, Chap. 9).

5.2 Musical Meaning and the Discourse of Timbre

Listening to a sound (speech, music, environmental events, etc.) involves not only detection-perception of the acoustic signal, but also the interpretation of auditory information (e.g., pitch or the lack thereof, timbre, duration, dynamics). According to Reybrouck (2013), musical semantics, the processing of meaning emerging from musical auditory information, relies on evolutionarily older mechanisms of meaningfully reacting to nonmusical sound, and

> "… listeners can be conceived as *adaptive* devices, which can build up new semiotic linkages with the sounding world. These linkages can be considered as by-products of both biological and cultural *evolution* and can be helpful in providing coordinative frameworks for achieving diversity of thought, cultural invention, social interaction and optimal coregulation of affect" (pp. 602–603; emphasis added).

Combining previous theoretical accounts of musical semantics with empirical neurobiological evidence, Koelsch (2011) concluded that there are three fundamentally different classes of musical meaning: *extramusical*, *intramusical*, and *musicogenic*. Extramusical meaning arises from the interpretation of musical sound cues through iconic, indexical, and symbolic sign qualities. Iconic qualities resemble qualities of objects and abstract concepts. Indexical meaning emerges from emotion and intention recognition. Symbolic meaning emerges from social and cultural associations. For example, a musical excerpt may sound buzzing, warm, complex, happy, ethnic, patriotic, and so on. Intramusical meaning emerges from the interpretation of structural references between musical units without extramusical associations, such as chord functions during the course of a cadence. Finally, musicogenic refers to meaning that stems from the interpretation of physical, emotional, and self-related responses evoked by musical cues, as opposed to interpreting musical cues per se. A musical performance can thus prompt one to dance, shed tears, or remember a past experience. Within the framework posited by Koelsch (2011), verbal attributes of timbral qualities can generally be thought of as falling into the class of iconic signs (Zacharakis et al. 2014).

5.2.1 Speaking about Sounds: Discourse Strategies

Wake and Asahi (1998) used musical, vocal, and environmental stimuli, and pairs of naïve listeners to study how they describe different types of sounds. Unlike sound experts (i.e., musicians, composers, sound artists, recording engineers, sound and music scholars) the naïve listeners lack a specialized auditory vocabulary. One person in each pair listened to a sound and subsequently described it to their interlocutor, who then had to imagine the described sound and, after listening to the actual stimulus, assess the similarity between the two. The verbalizations used to convey the different sounds were mainly of three types. The first type describes the *perception of the sound itself* using onomatopoeias (i.e., words or vocables considered by convention to phonetically mimic or suggest the sound to which they refer; e.g., chirin-chirin for the sound of a wind bell) or acoustic terminology (e.g., high pitched). The second type describes the *recognition of the sounding situation* using references to the object that made the sound (e.g., a bird) or the action that produced it (e.g., twittering) or other contextual information (e.g., in the morning). The third type describes the *sound impression* using metaphors and similes (e.g., clear, cool). Wake and Asahi (1998) proposed a model of auditory information processing, according to which recognition and impression are processed either independently (perception then recognition or impression) or sequentially (perception then recognition then impression).

In his empirical ethnographic research on the management of talk about sound between music professionals in the United States, Porcello (2004) identified five strategies that are common to the discourse of timbre among producers and engineers: (1) spoken/sung vocal imitations of timbral characteristics; (2) lexical onomatopoeic metaphors; (3) pure metaphor (i.e., non-onomatopoeic, generally referencing other sensory modalities or abstract concepts); (4) association (citing styles of music, musicians, producers, etc.); (5) evaluation (judgements of aesthetic and emotional value). Thus, a snare drum might sound like /dz:::/ and a muted trombone like wha-wha, a wolf tone on the cello (a persistent beating interaction between string vibrations and sympathetic body resonances) is usually howling and rough or harsh, and a violin tone might sound baroque or like Menuhin or beautiful. In comparison to the taxonomy of Wake and Asahi (1998), Porcello (2004) distinguishes between lexical onomatopoeias and vocal mimicry of nonvocal timbres, including in the latter category nonlexical onomatopoeias, and also considers three types of sound impression descriptions: pure metaphor, association, and evaluation.

Porcello (2004) further advances a distinction between vocal imitations and onomatopoeias on the one hand (which he calls "sonic iconicity") and the pure iconicity of metaphors originating in nonauditory sensory experiences or abstract concepts on the other hand. These, he observes, are usually "codified, especially among musicians and sound engineers," (Porcello 2004, p. 747). Following their investigation of the relation between verbal description and gestural control of piano timbre, Bernays and Traube (2009, p. 207) similarly concluded that "high level performers … have developed over the years of practice … an acute perceptive sensibility to slight sonic variations. This … results in an extensive vocabulary developed to describe the nuances a performer can detect." Furthermore, as noted by Traube (2004), this vocabulary is traditionally communicated from teacher to student in both the musician and sound engineer communities.

Lemaitre and colleagues (2010) analyzed free sortings of environmental sounds made by expert and nonexpert listeners along with scores of source-cause identification confidence and source-cause verbalizations. For the latter, participants were asked to provide nonmetaphorical nouns and verbs to describe the object and action that produced each sound. Participants were also asked to describe what sound properties they considered in grouping different sounds together. They showed that naïve listeners categorized environmental sounds primarily on the basis of source-cause properties. When these could not be identified, nonexpert listeners turned to the timbral properties of the sound, which they described using metaphors or vocal imitations. In contrast, musicians and other expert listeners relied more on timbral characteristics, verbalizing them using metaphors almost exclusively. This finding may offer support to the auditory information processing model proposed by Wake and Asahi (1998), who assert that timbral impression is processed independently of or following source recognition. It could also help to explain why Porcello's taxonomy of timbre verbalizations, which is derived from the discourse of sound experts, does not include descriptions of the physical cause of a sound, such as those grouped under "sounding situation" by Wake and Asahi (whose taxonomy is based on verbalizations by nonexpert listeners).

Wallmark (2018) conducted a corpus linguistic analysis of verbal descriptions of instrumental timbre across eleven orchestration treatises. The collected verbalizations were categorized according to: (1) affect (emotion and aesthetics); (2) matter (physical weight, size, shape); (3) crossmodal correspondence (borrowed from other senses); (4) mimesis (sonic resemblance); (5) action (physical action, movement); (6) acoustics (auditory terminology); and (7) onomatopoeia (phonetic resemblance). This scheme is very similar to the one suggested by Porcello (2004), whose notion of "pure" metaphor could be seen as encompassing categories (2) to (6). Whereas onomatopoeic words were prevalent among music producers and engineers in Porcello's study, they accounted for a mere 2% of Wallmark's orchestration corpus, driven primarily by a small number of mostly percussion instruments. In fact, certain instruments and instrument families were found to have a systematic effect on verbal description category. For example, the trombone was described more frequently with affect and mimesis than other brass instruments, while the violin, viola, and cello all shared similar descriptive profiles (cf., Saitis et al. 2017). By means of principal components analysis, the seven categories were further reduced to three latent dimensions of musical timbre conceptualization: *material* (loaded positively onto onomatopoeia and matter), *sensory* (crossmodal and acoustics), and *activity* (action and mimesis).

Notwithstanding the diverse metaphorical timbre lexicon in orchestration books, taxonomies of musical instruments and the kinds of sounds they produce are usually based on the nature of the sound-producing material and mechanism. Koechlin (1954–1959; cited in Chiasson et al. 2017, p. 113–114) proposed instead to organize instrument sounds for orchestration purposes on the basis of volume and intensity. Volume is described as an impression of how much space an instrument sound occupies in the auditory scene ("extensity" is used by Chiasson et al. 2017; see also Rich 1916). Based on an inverse relationship between volume and intensity, Koechlin (cited in Chiasson et al. 2017) further proposed a third attribute of density versus transparency: a musical sound is dense when it is loud but with a small volume, and it is transparent when it has a large volume but low intensity. There is evidence that in the later Middle Ages it was typical to think of musical instruments in terms of volume of sound (Bowles 1954). In orchestras, and for other musical events, instruments with a big, loud sound (*haut* in French) would be grouped together against those with a small, soft sound (*bas*).

Schaeffer (1966) offered a typo-morphology of "sonorous objects" (i.e., sounds experienced by attending to their intrinsic acoustic properties and not to their physical cause) based on sustainment (*facture* in French) and mass. Sustainment refers to the overall envelope of the sound and mass is described as "the quality through which sound installs itself … in the pitch field" (Schaeffer 1966, p. 412), which appears similar to Koechlin's notion of volume. Interestingly, Koechlin and

Schaeffer were both French, shared a composition background, and published their typologies within 10 years of each other. Mass extends the concept of pitch in pure tones (i.e., single frequencies) and tonal sounds (i.e., nonnoisy) to include sounds with fluctuating or indeterminate pitch (e.g., cymbals, white noise). Each mass has a particular timbre associated with it—a set of "secondary" qualities that are either nonexistent (pure tones) or exist at varying degrees from being dissociated (musical notes) to indistinguishable (white noise) from mass. Given the definition of sonorous objects, Schaeffer's timbre is free from any source-cause associations and is thus situated clearly in the realm of quality as opposed to identity (Siedenburg, Saitis, and McAdams, Chap. 1).

In tonal sounds, Schaeffer argues, mass can be low or high (in terms of location in the pitch field) and thick or thin (in terms of extensity in the pitch field); timbre can be dark or light (location), ample or narrow (extensity), and rich or poor (in relation to the intensity of the mass). The latter appears closely related to Koechlin's notion of density as they both describe a mass or volume, respectively, in relation to its intensity. In Smalley's (1997) *Theory of Spectromorphology*, which has its origins in Schaeffer's ideas, pitch field is replaced by "spectral space". The latter is described in terms of emptiness versus plenitude (whether sound occupies the whole space or smaller regions) and of diffuseness versus concentration (whether sound is spread throughout the space or concentrated in smaller regions). Like Koechlin and Schaeffer, Smalley also relies on extra-auditory concepts to serve as discourse for an organization of auditory material that focuses on *intrinsic* features of the sound independently of its source.

5.2.2 Metaphors We Listen With

Wallmark (2014) argues that the metaphorical description of timbre is not simply a matter of linguistic convention, and what Porcello singles out as "pure metaphor" is central to the process of conceptualizing timbre by allowing the listener to communicate subtle acoustic variations in terms of other more commonly shared sensory experiences (nonauditory or auditory-onomatopoeic) and abstract concepts. De Ceuster (2016) points out that timbre has been described with metaphors based on experiences since the presumed birth of the term in the mid-eighteenth century (Dolan 2013). Jean-Jacques Rousseau's "Tymbre" entry in Diderot and D'Alembert's *Encyclopedié* reads:

> A sound's *tymbre* describes its harshness or softness, its dullness or brightness. Soft sounds, like those of a flute, ordinarily have little harshness; bright sounds are often harsh, like those of the *vielle* [medieval ancestor to the modern violin] or the oboe. There are even instruments, such as the harpsichord, which are both dull and harsh at the same time; this is the worst *tymbre*. The beautiful *tymbre* is that which combines softness with brightness of sound; the violin is an example (cited and translated in Dolan 2013, p. 56).

Building on accounts of ecological and embodied cognition, Wallmark (2014) proposes an embodied theory of timbre whereby metaphorical descriptions are indexes of conceptual representations grounded in perception and action. They can be grouped into three categories based on the conceptual metaphors (Lakoff and Johnson 2003): (1) *instruments are voices* (e.g., nasal, howling, open); (2) *sound is material* (e.g., bell-like, metallic, hollow, velvety); and (3) *noise is friction* (e.g., harsh, rough) (cf., Wallmark 2018). The *sound is material* metaphor can be broken down into four subtypes: (2a) naming the source directly (e.g., a bell-like sound); (2b) referencing the physical qualities of the source (e.g., a metallic sounding cymbal); (2c) blending physical and connotative elements of source and sound (e.g., a hollow bassoon); and (2d) referencing physical qualities of unrelated objects (e.g., velvety strings).

Why are instruments voices? Consider phonemes. They can be categorized based on distinctive features associated with the physiology of voice production and articulation that are generally inherent in all languages (Jakobson and Halle 1971). Phonemes can be nasal (coupling between the oral and nasal cavities) or oral (no coupling); compact (spectral dominance of a single central formant when the mouth is wide open) versus diffuse; strident (airstream forced to strike the teeth, high-intensity fricative noise) versus mellow; tense or lax (greater versus lesser deformation of the vocal tract); grave (larger and less compartmented mouth cavity volume, concentration of energy in the lower register) versus acute; flat (smaller lip opening but larger between-lip area, weakening of upper frequencies) or nonflat; and sharp (dilated pharyngeal pass, strengthening of upper frequencies) versus nonsharp. In singing, a low versus high laryngeal position produces a covered versus open vocal timbre or simply a low versus high pitch (Miller 1986). In medicine, hoarse is used to describe the presence of high frequency noise components accompanied by decreased harmonics in the voice due to laryngeal diseases (Isshiki et al. 1969). Attributes such as howling, throaty, hissing, and breathy eventually refer to the associated vocal source or as Sundberg (2013, p. 88) puts it: “The perception of voice seems to be influenced by familiarity with one’s own voice production.” This observation echoes the motor theory of speech perception, which considers that the latter is based on articulatory motor representations (Liberman and Mattingly 1985) and which Wallmark (2014) extends to a motor theory of all timbre perception in preparation for the *instruments are voices* metaphor.

Albersheim (1939) drew analogies between vowels and colors to propose a geometrical model of *acoustic color* (*Akustischer Farbenkörper* in German) in the form of a cylinder. Its height and radius represented variation in color brightness and saturation, respectively. Changes in color hue were mapped onto a helical line along the surface of the cylinder. Slawson (1985) developed a theory of *sound color*, which he defined as the static spectral envelope of a sound, as opposed to its temporally varied spectrum, based on the distinctive phoneme features of *openness*, *acuteness*, and *laxness*, and their relation to the pitch-invariant formant structure of vowels. The term “openness” was chosen as a perceptually more intuitive depiction of compactness. More open vowels have a higher first formant, while acuteness increases with increasing frequency of the second resonance. Lax vowels have a lower total energy that is less spread out over the spectrum. A fourth dimension was termed *smallness*: the lower the first and second formants are, the smaller the vowel. Schumann (1929), Reuter (1997), and Lembke and McAdams (2015), among others, have discussed the vowel-like pitch-invariant formant structure of many (but not all) musical instruments and its role in timbre perception.

In other words, timbre can be experienced with reference to the human and non-human voice—a conceptualization already evident in Helmholtz's (1877) choice to synthesize vowel-like sounds for his *Klangfarbe* experiments and in Schilling's definition of the German term as "denoting mostly the accidental properties of a voice" (Schilling 1840, p. 647; cited in Kursell 2013). Timbre can also be experienced as a material object that can be seen, touched, and even tasted. Furthermore, noise-like timbres (e.g., excessive high-frequency content, inharmonicity, flat spectrum) can be understood in terms of frictional material interaction. Very similar metaphorical conceptualizations can be found in verbalizations of other perceptual aspects of sound, such as pitch and loudness (Eitan and Rothschild 2011; Saitis et al. 2017). In general, conceptual metaphors of timbre and auditory semantics may originate in more universal neural processes and structures beyond auditory cognition (cf., Gallese and Lakoff 2005; Walsh 2013).

5.3 Semantic Spaces of Timbre

Scientific interest in timbre semantics started as early as the experimental exploration of timbre itself (Helmholtz 1877; Stumpf 1890). Stumpf (1890) proposed that the various verbal attributes of timbre can be summarized on the basis of semantic proximities by three pairs of opposites: dark–bright (*dunkel–hell* in German), soft–rough (*weich–rauch*), and full–empty (*voll–leer*). Hereafter, these symbols will be used: '–'to indicate antonyms and '/' to indicate synonyms. Discussing a set of psychoacoustic experiments, Lichte (1941) concluded that brightness, roughness, and fullness, as defined by Helmholtz, form independent attributes of sound in addition to pitch and loudness. More systematic efforts to understand the complex multivariate character of timbre semantics were made possible by methodological tools such as factor analysis of ratings on verbal scales that were developed in the 1950s and were first applied to timbre by Solomon (1958) (Sect. 5.3.1). Studies using multidimensional scaling of adjective dissimilarities and psycholinguistic analyses of verbalization tasks have provided additional insight regarding particular aspects that contribute to the semantic description of timbre (Sect. 5.3.2).

5.3.1 Semantic Scales: Methodology and Main Results

Osgood (1952) developed a quantitative method for measuring meaning based on the use of multiple verbal scales. Each scale was defined by pairs of antonymic descriptive adjectives, such as dark–bright and smooth–rough, which he termed *semantic differentials*. The method postulates a semantic space within which the operational meaning of a given concept can be specified. This "space" is physically thought of as a Euclidean spatial configuration of unknown dimensionality; each semantic differential represents an experiential continuum, a straight line function that passes through the origin of this space. Many different continua are psychologically equivalent and, hence, may be represented by a single latent dimension. The minimum number of such (orthogonal) dimensions can be recovered by means of factor analysis and used to define the semantic space of the concept. Ratings on semantic scales also can be analyzed with principal component analysis or, when appropriately reorganized (e.g., dissimilarity distances, cross-correlations), clustering or multidimensional scaling techniques. The reliability and validity of the semantic differential model depend on a number of methodological considerations (Susini et al. 2012; Saitis et al. 2015). For example, it is important to use verbal scales that are psychologically relevant and commonly interpreted across all raters. And even then, the derived factors are not always easy to interpret with respect to the scales and/or raters.

Solomon (1958) had sonar technicians rate recordings of passive sonar sounds on bipolar scales comprising perceptual attributes (e.g., smooth–rough) that are typically used by experienced sonar operators but also a large number of aesthetic–evaluative adjectives (e.g., beautiful–ugly). A seven-factor solution was obtained, which accounted for only 42% of the total variance in the collected ratings. The first and most salient factor (15%) indicated a "magnitude" dimension, explained by such scales as heavy–light and large–small. The third factor (6%) was identified by such words as clear, definite, and obvious, and labeled as "clarity." The remaining factors were essentially aesthetic–evaluative, probably because many such differentials were used in the design of the study. Generally speaking, such scales are likely to be of little help when one tries to access perceptual representations through language, as affective reactions tend to be less stable across individuals than sensory descriptions.

Jost (1967; cited in Webster et al. 1970, p. 481–483) carried out a semantic differential study of four clarinet notes played at six different loudness levels and found two salient factors of density and volume. However, these appeared to correlate with stimuli variations in pitch and loudness, respectively. Von Bismarck (1974a) sought to address three important issues with applying semantic differentials to the study of timbre semantics: selecting verbal attributes that are perceptually relevant, normalizing sound stimuli for pitch and loudness, and psychophysically explaining the extracted factors. Sound stimuli comprised synthetic steady-state signals of two types: vowel-like and instrument-like harmonic complexes, and consonant-like noises. These had spectral envelopes varying systematically along three parameters: frequency location of overall energy concentration, slope of the envelope, and frequency location of energy concentrations within the spectrum. All sounds were normalized in loudness by means of perceptual adjustment to a given reference. The harmonic complexes were further equalized in fundamental frequency at 200 Hz. Sixty-nine differential scales were initially rated for suitability to describe timbre on a scale from "very unsuitable" to "highly suitable". From thirty-five scales with the highest mean suitability ratings, seven scales deemed synonymous were further discarded. The scales soft–loud and low–high were included to test the effectiveness of loudness and pitch normalization, respectively.

Factor analysis of ratings by a group of musicians and another group of nonmusicians yielded similar, although not identical, four-factor solutions that explained more than 80% of the variance in the data. The four factors were defined by the differentials dull–sharp, compact–scattered, full–empty, and colorful–colorless. Although participants were instructed to ignore pitch and loudness as much as possible, ratings on the soft–loud and low–high scales were highly correlated with those on dull–sharp and dark–bright, respectively. This illustrates how the same word can have different connotations in different contexts. Even when sounds were equalized in loudness and pitch, listeners still used related attributes to describe other impressions. In agreement with the view that verbal attributes of timbre are "codified" among musically trained listeners (see Sect. 5.2.1), ratings from nonmusicians were more scattered than those of musicians. Prompted by the finding that the dull–sharp factor explained almost half of the total variance in the data, von Bismarck (1974b) confirmed in subsequent psychoacoustic experiments that a dull–sharp scale had desirable measurement properties (e.g., doubling, halving) and concluded that sharpness may represent an attribute of sounds distinguishable from pitch and loudness.

Von Bismarck's is arguably the first comprehensive investigation of timbre semantics, markedly improving upon the earlier studies, but certain aspects have been questioned. For example, aesthetic-evaluative and affective scales were still used. In addition, the preliminary assessment of whether or not a scale was suitable for describing timbre was carried out in an undefined context, without presentation of the timbres to be described, while further discarding of scales was based on an arbitrary judgement of word synonymy. Perhaps more importantly, a *semantic* issue with the semantic differentials is the assumption of bipolarity that underlies the model (Heise 1969; Susini et al. 2012). Are soft–loud and dark–bright always true semantic contrasts? Is sharp the true semantic opposite of dull when talking about timbre?

One way to address potential biases associated with prescribing antonymic relationships between adjectives is to use adjective checklists. These were used extensively in musical affect research up until the late 1950s (for a review, see Radocy and Boyle 2012) but have largely been replaced by semantic scales. Similarly to von Bismarck (1974a), Pratt and Doak (1976) attempted to first find verbal scales suitable for describing timbre. An initial list of 19 "commonly used" adjectives was reduced to seven items by means of a checklist task. By (arbitrarily) discarding synonyms and "not very useful" words, the list was further reduced to the attributes brilliant, rich, and warm; dull, pure, and cold, respectively, were (arbitrarily) chosen as opposites to form semantic differentials. From ratings of different synthesized harmonic spectra on the three scales, it was found that the former were most consistently discriminated by the brilliant–dull scale.

In a separate study (Abeles 1979), each of twenty-four recorded isolated clarinet notes was presented three times, each time with five adjectives randomly selected from a list of forty words. Three independent groups of clarinetists, nonclarinetist musicians, and nonmusicians were asked to check as many adjectives as they thought best described the timbre of each note. Factor analysis of the combined data across the three listener groups (no individual group analyses were reported) yielded a three-factor solution of shape (round/centered–pinched/thin), density (clear/brilliant–fuzzy/airy), and depth (resonant/rich/projecting; no negatively loaded adjectives were reported). Edwards (1978) and Pratt and Bowsher (1978) found a very similar set of semantic dimensions for the trombone (see Sect. 5.3.2), which is also a wind instrument.

Kendall and Carterette (1993a, b) attempted a systematic use of verbal scales bounded by an attribute (e.g., bright) and its negation (e.g., not bright), which was termed the verbal attribute magnitude estimation (VAME) method because the task for the rater is to assess how much of a single attribute is possessed by a stimulus. Unipolar scales offer a way of dealing with polysemy and nonexact antonymy within the semantic differential framework. Accordingly, antonymic or synonymic relationships can be assessed a posteriori through negative or positive correlations between ratings on different unipolar scales.

A first pair of experiments (Kendall and Carterette 1993a) sought to explore the extent to which von Bismarck's (1974a) semantic space, which had resulted from synthetic vowel-like sounds, is relevant in describing the timbre of natural (recorded) instrument sounds. The stimuli comprised dyads of wind instrument notes produced in unison, and they were rated by nonmusicians on eight VAME scales that loaded high on the first (hard, sharp, loud, complex), second (compact, pure), and fourth (dim, heavy) von Bismarck factors. Analyses converged to a two-dimensional solution accounting for almost 98% of the variance; however, it mapped weakly onto a two-dimensional similarity space of the same dyads, prompting the authors to conclude that von Bismarck's scales were less relevant in rating natural versus synthetic timbres. In subsequent experiments (Kendall and Carterette 1993b), the same stimuli were rated by musicians on twenty-one VAME scales induced from adjectives describing instrumental timbre in an orchestration book. Similar analyses resulted in a two-dimensional semantic space of nasal–rich and brilliant–reedy adjectives, which explained 96% of the data variance and corresponded more strongly with similarity ratings.

The work of Kendall and Carterette constitutes the first systematic effort to combine semantic ratings with similarity judgements to directly examine the relationship between the perception of timbre and its verbal communication. In this context, these results illustrate that the validity of a semantic space as a perceptual construct depends on a number of issues such as the type of sounds tested, the type of verbal scales used, and the musical background of raters. Especially when considering differences in how musically experienced versus naïve listeners conceptualize timbral qualities (see Sect. 5.2.1), it is plausible that the better results obtained in the second set of experiments (Kendall and Carterette 1993b) were not only a result of selecting more relevant semantic scales but also of recruiting musically trained listeners. Von Bismarck (1974a) and Abeles (1979) both found that in rating the same sounds on the same semantic scales musicians were generally more consistent than nonmusicians.

The nasal–rich dimension of Kendall and Carterette (1993b) summarizes descriptions of nasal/edgy/brittle/weak/light versus rich/round/strong/full. It thus appears to correspond to the shape factor found by Abeles (1979) for clarinet sounds. Abele's density factor seems to be closer to Kendall and Carterette's brilliant–reedy dimension, which relates to impressions of brilliant/crisp/pure versus reedy/fused/warm/complex. In some agreement with these two studies, Nykänen et al. (2009) found four semantic dimensions for a set of saxophone notes, namely, warm/soft, back vowel-like sounding, sharp/rough, and front vowel-like sounding. Considering that back versus front vowels tend to be perceived as dark/round versus bright/thin (Jakobson and Halle 1971), two broader dimensions alluding to shape (warm/soft–sharp/rough) and density (dark/round–bright/thin) may be hypothesized. It therefore appears that most wind instrument timbres can be positioned within a common semantic space. How does this space adapt when sounds from other instrument families are included? Kendall et al. (1999) found that adding a violin note did not affect the semantic space; however, its mapping onto the corresponding perceptual space was less robust.

Using fifteen VAME scales, Disley et al. (2006) obtained a four-dimensional semantic space for twelve orchestral instrument notes of same pitch: bright/thin/harsh/clear–dull/warm/gentle/rich, pure/percussive/ringing–nasal, metallic–wooden, and evolving. Ratings remained fairly consistent across multiple repetitions. Several listeners noted that they used metallic and wooden to describe the recognized material of the instrument rather than a timbral quality, which would explain the loading of these scales on a separate component (one could expect metallic to correlate with bright/harsh and wooden with warm/rich). Similarly, the presence of a fourth dimension solely defined by evolving is likely due to reported listener difficulties in understanding what it meant, although the moderate loading of rich on the same component might indicate a spectral flux type of dimension (see Sect. 5.4.1).

Using a more diverse set of stimuli (twenty-three isolated notes from acoustic, electromechanical, and electronic instruments, with different pitches), twice as many VAME scales, and analyses that accounted for nonlinear relationships between the semantic variables, Zacharakis et al. (2014) arrived at a three-dimensional space summarized as luminance (brilliant/sharp–deep), texture (soft/rounded/warm–rough/harsh), and mass (dense/rich/full/thick–light). This space was largely similar across two independent groups of native English and Greek-speaking listeners (musically experienced). Two different groups of English and Greek listeners provided dissimilarity ratings of the same set of sounds and the respective three-dimensional spaces derived from multidimensional scaling (MDS) were also found to be highly similar. Comparisons between the semantic and perceptual spaces illustrated strong correlations of luminance and texture, on the one hand, and texture with two of the three MDS dimensions on the other, independent of native language. Texture appeared to contribute to all three MDS dimensions. Results for mass were less conclusive. Moderately similar results have been obtained for an even larger set of musical sounds (forty-two sustained orchestral instrument notes of the same pitch) using bipolar scales and different methods of analysis (Elliott et al. 2013). A strong, but not one-to-one, correspondence between semantic and perceptual dimensions of timbre had previously been shown by Samoylenko et al. (1996) and Faure (2000), who collected free verbalizations during dissimilarity ratings.

5.3.2 Further Findings from Verbalization and Verbal Dissimilarity Tasks

Verbalization tasks, where participants are asked to describe timbral impressions in their own words, offer an alternative means of exploring the semantics of timbre. They can be used as a standalone method (Traube 2004; Saitis et al. 2017), or to complement a preceding task (e.g., describe timbral differences during pairwise sound comparisons: Samoylenko et al. 1996; Faure 2000), or to help design further experiments (e.g., extract relevant adjectives for anchoring semantic scales: Rioux and Västfjäll 2001; Grill 2012). Verbalization can be *free*, in the sense that very general open-ended questions are asked and no restriction is imposed on the format of the response, or *constrained*, where questions are more structured and responses must conform to a certain format. A qualitative method of deriving semantic proximities from verbalization data relies on theoretical assumptions about cognitive categories and their relation to natural language (Dubois 2000). From what is being said and how it is being said, relevant inferences can be derived about how people conceptualize sensory experiences (semantic level) and can be further correlated with physical parameters (perceptual level).

Traube (2004) asked classical guitar players to freely describe the timbre of their instrument in relation to how it is produced. The ten most commonly used adjectives were dry, nasal, thin, metallic, bright, round, warm, thick, velvety, dark. By combining linguistic analysis and acoustic measurements, a strong correspondence was found between the plucking position along the string, the frequency location of the generated comb filter formants, and the use of adjectives describing vowel-like timbre for similarly located vocal tract formants, which echoes the *instruments are voices* metaphor (Sect. 5.2.2). As an example, adding the nasal and oral cavities (nasal voice) causes a broadening of all vocal tract formant bandwidths and a flattening of spectral peaks in the range 300–2500 Hz (Jakobson and Halle 1971; Mores 2011). Traube found that guitars sound more nasal/bright/dry when plucked closer to the bridge because of analogous spectral effects. Conversely, plucking between the sound hole and the fingerboard produces spectra similar to nonnasal vowels and is perceived as more velvety/dark/round.

Rioux and Västfjäll (2001) and Saitis et al. (2017) have provided further evidence that, while perceived variations in how an instrument sounds rely on variations in style and the expertise of different musicians (Saitis et al. 2012), the broader semantic categories emerging from verbal descriptions remain common across diverse musical profiles, thus reflecting a shared perception of acoustic information patterns. Importantly, the verbal data revealed that vibrations from the violin body and the bowed string (via the bow) are used as extra-auditory cues that not only help to better control the played sound but also contribute to its perceived qualities. For example, recent research on the evaluation of piano and violin quality has revealed that an increase in the vibrations felt at the fingertips of pianists and the left hand of violinists can lead to an increase in perceived sound loudness and richness (Saitis et al. 2018). Also, impressions like bright and rich mostly refer to the sustained part of a note, while words like soft tend to describe qualities of transients (cf., Brent 2010; Bell 2015).

An example of constrained verbalization is the *repertory grid technique*. Listeners form bipolar constructs (i.e., antonymic pairs of adjectives) by articulating the difference between two sounds taken from a larger pool that is relevant to the aims of the task at hand (referred to as elements). Alternatively, three sounds are presented and listeners are first invited to select the least similar one and subsequently to verbally explain their grouping. Finally, listeners are asked to rate all elements on each new construct. The resulting grid of constructs and elements, essentially semantic differential ratings, can then be evaluated with factor analytical, clustering, or multidimensional scaling techniques. Using this method, Grill (2012) found an expanded semantic space for electroacoustic "textures", which combined dimensions pertinent mostly to such sounds (ordered–chaotic or coherent–erratic, homogeneous–heterogeneous or uniform–differentiated) with dimensions commonly found for voices and instruments (high–low or bright–dull, smooth–coarse or soft–raspy, tonal–noisy).

A semantic space can also be derived quantitatively through MDS of pairwise distances in a list of adjectives. Moravec and Štěpánek (2003) initially asked conductors, composers, engineers, teachers, and musicians (three groups of bowed-string, wind, and keyboard performers) to provide words they typically use to describe the timbre of any musical instrument. The four most frequently mentioned words across all respondents (sharp, gloomy, soft, clear) were also among the four most frequently used in each of the three musician groups. Still, some within-group preferences were observed. Bowed-string players used sweet and warm more frequently than both keyboard and wind performers. Similarly, narrow was much more popular with wind musicians. The thirty most frequently reported adjectives were subjected to dissimilarity ratings (Moravec and Štěpánek 2005) and MDS identified three dimensions closely matching luminance, texture, and mass (Zacharakis et al. 2014), namely, gloomy/dark–clear/bright, harsh/rough–delicate, and full/wide–narrow, respectively.

Edwards (1978) collected a corpus of free verbalizations of trombone sound quality through interviews and a postal survey of over 300 trombone performers. A subset of the verbal data was arranged in terms of semantic similarity by the author himself on the basis of proximities identified in the corpus. This kind of dissimilarity matrix was subsequently subjected to MDS. With respect to timbre, two dimensions of small–wide and dull/round–clear/square emerged. A different subset of the verbalizations indicated a third timbral aspect referring to "amount" and "carrying" or "penetrating" properties of sound. These seem to generally agree with the findings of Abeles (1979), Kendall and Carterette (1993b), and Nykänen et al. (2009).

In another trombone study, Pratt and Bowsher (1978) selected the scales compact–scattered, dull–bright, and not penetrating–penetrating to correspond to Edwards' three dimensions. It was found that the second and third scales were good discriminators of trombone timbres but compact–scattered was not. Indeed, the latter may not indicate size, which is the label Edwards gave to his first dimension, but may indicate density (see Sect. 5.2.1).

Fritz et al. (2012) had violinists arrange sixty-one adjectives for violin timbre on a two-dimensional grid (EXCEL), so that words with similar meanings lay close together and those with different meanings lay far apart. The collected grids were converted into dissimilarity matrices using a custom distance metric between two cells (see p. 793 in Fritz et al. 2012) and MDS yielded three dimensions: warm/rich/mellow versus metallic/cold/harsh (richness; texture), bright/responsive/lively versus muted/dull/dead (resonance; projection), and even/soft/light versus brash/rough/raspy (texture; clarity). The parenthetical terms potentially correspond to semantic categories from the cognitive model proposed by Saitis et al. (2017). In both studies, violinists used words like lively, responsive, ringing, and even bright to describe the "amount of sound" perceived "under the ear" (resonance) and in relation to spatial attributes (projection). Differences between the labels of the found semantic dimensions for trombone (wind) and violin (bowed string) timbre seem to generally agree with those observed by Moravec and Štěpánek (2003).

In the piano study of Bernays and Traube (2011), fourteen adjectives extracted from spontaneous verbalizations yielded a four-dimensional MDS space. Based on the first two dimensions (78% of the total variance explained) and additional hierarchical clustering, five adjectives were proposed to best represent a semantic space for piano timbre: bright, dry, dark, round, and velvety. Lavoie (2013) performed MDS on dissimilarities between adjectives describing classical guitar timbre. In agreement with Traube (2004), a dimension of velvety/dark–bright/dry was obtained, related to whether the string is plucked between the sound hole and the fingerboard versus closer to the bridge (like nasal); a dimension of round/bright–dull/thin was associated with sound resonance and projection. It is worth noting the highly similar labels of the reported semantic spaces across the two instruments. To a certain extent, this may reflect shared conceptualization structures between musicians whose primary instrument produces impulsive string sounds. On the other hand, given that all three studies were conducted with musicians from the Montreal region, it may be that these results mirror a verbal tradition specific to that geographic location, possibly due to a strong influence by one or more particular teachers in the area (cf., Saitis et al. 2017).

5.4 Semantic Spaces of Timbre Revisited

Despite important methodological differences, the findings described in the previous section show remarkable similarities when certain classes of timbres (e.g., individual instrument families) and mixed sets across distinct classes (e.g., various orchestral instruments) are rated on verbal scales, but similarities are also evident when verbal descriptions are collected in the absence of sound examples (e.g., verbalization tasks, adjective dissimilarity ratings). The most salient dimensions can be interpreted broadly in terms of brightness/sharpness (or luminance), roughness/harshness (or texture), and fullness/richness (or mass). The boundaries between these dimensions are sometimes blurred, while different types of timbres or scenarios of timbre perception evoke semantic dimensions that are specific to each case (e.g., nasality, resonance/projection, tonalness–noisiness, compact–scattered). Generally, no striking differences between expert and naïve listeners are observed in terms of semantic dimensions, although the former tend to be more consistent in their perceptions than the latter. In this section, the identified semantic dimensions of timbre are examined further through looking at their acoustic correlates (Sect. 5.4.1) and comparisons between different languages and cultures (Sect. 5.4.2).

5.4.1 Acoustic Correlates

Impressions of brightness in timbre perception are typically found correlated with the spectral centroid, a scalar descriptor defined as the amplitude-weighted mean frequency of the spectrum (Siedenburg, Saitis, and McAdams, Chap. 1; Caetano, Saitis, and Siedenburg, Chap. 11), which indicates the midpoint of the spectral energy distribution (cf., Lichte 1941). In other words, frequency shifts in spectral envelope maxima are systematically perceived as changes in brightness. The spectral centroid is typically found correlated with one of the dimensions (usually of three) that describe timbre dissimilarity spaces. A higher proportion of high-frequency energy also characterizes brightness in timbral mixtures arising from multitrack recorded music, although the absence of high pitch in such stimuli rendered them as less bright (Alluri and Toiviainen 2010). This is because frequency shifts in pitch, too, are systematically perceived as changes in brightness (Cousineau et al. 2014; Walker 2016). The sharpness factor in von Bismark's (1974a) study (dull–sharp, soft–hard, dark–bright) was also strongly related to the frequency position of the overall energy concentration of the spectrum, with sharper/harder/brighter sounds having more energy in higher frequency bands. Similarly, Bloothooft and Plomp (1988) observed that verbal attributes of stationary sung vowels related to sharpness (including sharp–dull, shrill–deep, metallic–velvety, angular–round, and cold–warm) referred primarily to differences in spectral slope between the vowels. Acute (i.e., sharp) phonemes are also characterized by a concentration of energy in the higher frequencies of the spectrum (Jakobson and Halle 1971; Slawson 1985).

A model for estimating sharpness, originally proposed by von Bismarck (1974b), calculates the midpoint of the weighted specific loudness values in critical bands (Fastl and Zwicker 2007). Critical bands correspond to equal distances along the basilar membrane and represent the frequency bands into which the acoustic signal is divided by the cochlea. Grill (2012) found a strong correlation between bright–dull electroacoustic textural sounds and the sharpness model, which is consistent with the origin of the latter in psychoacoustic experiments with wideband noise spectra. However, Almeida et al. (2017) showed that the sharpness model insufficiently predicted brightness scaling data for tonal sounds. Marozeau and de Cheveigné (2007) proposed a spectral centroid formula based on the same concept of weighted partial loudness in critical bands, which better modeled the brightness dimension of dissimilarity ratings and was less sensitive to pitch variation compared to the classic spectral centroid descriptor.

Yet another verbal attribute that has been associated with spectral energy distribution is nasality. Etymologically, nasality describes the kind of vocal sound that results from coupling the oral and nasal cavities (Sects. 5.2.2 and 5.3.2). However, it is sometimes used to describe the reinforcement of energy in higher frequencies at the expense of lower partials (Garnier et al. 2007; Mores 2011). In violin acoustics, nasality is generally associated with a strong frequency response in the vicinity of 1.5 kHz (Fritz et al. 2012). Kendall and Carterette (1993b) found that nasal versus rich wind instrument sounds had more energy versus less energy, respectively, in the upper harmonics, with rich timbres combining a low spectral centroid with increased variations of the spectrum over time. Sounds with a high versus a low spectral centroid and spectral variation were perceived as reedy versus brilliant, respectively. Adding a violin note in a set of wind instrument timbres confirmed a strong link between nasality and the spectral centroid, but rich and brilliant were correlated only with spectral variation and only to some modest degree (Kendall et al. 1999). Helmholtz (1877) had originally associated the nasality percept specifically with increased energy in odd numbered upper harmonics, but this hypothesis remains unexplored.

Are timbral brightness and sharpness the same percept? Both of them relate to spectral distribution of energy, and most of the related studies seem to suggest at least partial similarities, but there is still no definite answer to this question. Štěpánek (2006) suggested that a sharp timbre is one that is both bright and rough. However, semantic studies of percussive timbre reveal two independent dimensions of brightness and sharpness/hardness (Brent 2010; Bell 2015). Brighter percussive timbres appear associated with higher spectral centroid values during attack time, while sharp/hard relates to attack time itself (i.e., sharper/harder percussive sounds feature shorter attacks). Attack time refers to the time needed by spectral components to stabilize into nearly periodic oscillations, and it is known to perceptually distinguish impulsive from sustained sounds (McAdams, Chap. 2). Furthermore, concerning brightness, there seems to exist a certain amount of interdependency with fullness. Sounds that are described as thick, dense, or rich are also described as deep or less bright and brilliant, while nasality combines high-frequency energy with low spectral spread and variability. The acoustic analyses of Marozeau and de Cheveigné (2007) and Zacharakis et al. (2015) suggest that brightness may not only relate to spectral energy distribution but also to spectral detail.

To further complicate things, a number of studies based on verbalizations that were collected either directly from musicians or through books and magazines of music revealed a semantic dimension of timbre associated with a resonant and ringing but also bright and brilliant sound that can project (Sect. 5.3.2). This suggests an aspect of timbre that is primarily relevant to playing an instrument and is associated with assessing how well its sound is transmitted across the performance space. It also suggests an interaction between perceived sound strength and timbral brightness. Based on sound power measurements and audio content analysis of single notes recorded at pianissimo and fortissimo across a large set of standard orchestral instruments (including some of their baroque and classical precursors), Weinzierl et al. (2018b) were able to show that the intended dynamic strength of an instrument can be identified as reliably by sound power as by combining several dimensions of timbral information. Indeed, the most important timbral cue in this context was found to be spectral skewness (Caetano, Saitis, and Siedenburg, Chap. 11) with a left-skewed spectral shape (i.e., a shift of the peak energy distribution toward higher frequencies) indicating high dynamic strength.

Helmholtz (1877) claimed that the sensation of roughness arises from the increasingly dissonant (unpleasant) sounding intervals formed between higher adjacent partials above the sixth harmonic. Empirical data from Lichte (1941) and later Schneider (1997) support this view, which has also lent itself to theories of musical tension (McAdams, Chap. 8). However, Stumpf (1898) disagreed with Helmholtz and provided examples of dissonant chords that were judged as not rough, highlighting a difference between *musical* dissonance and *sensory* dissonance. More recent evidence also suggests that roughness (expressing sensory dissonance) and musical dissonance may constitute distinct percepts (McDermott et al. 2010; Bowling et al. 2018). Physiologically, impressions of roughness and/or sensory dissonance can be linked to the inability of the cochlea to resolve frequency pair inputs whose interval is smaller than the critical band, causing a periodic “tickling” of the basilar membrane (Helmholtz 1877; Vassilakis and Kendall 2010).

Further psychophysical experiments have linked roughness to envelope fluctuations within a critical band produced by amplitude-modulation frequencies in the region of about 15–300 Hz (Fastl and Zwicker 2007; Vassilakis and Kendall 2010). For a given amplitude spectrum and a given modulation depth, modulations with an abrupt rise and a slow decay have been shown to produce more roughness than modulations with a slow rise and an abrupt decay (Pressnitzer and McAdams 1999). For electroacoustic sounds, the effect of sudden changes in loudness over broad frequency ranges is described as coarse and raspy (Grill 2012). Existing psychoacoustic models estimate roughness using excitation envelopes (Daniel and Weber 1997) or excitation-level differences (Fastl and Zwicker 2007) produced by amplitude modulation in critical bands. Nykänen et al. (2009) found that both models of sharpness (von Bismarck 1974b) and roughness (Daniel and Weber 1997) contributed to predictions of roughness of saxophone sound, but sharpness was a much more important contributor. However, and as noted already, these models were originally designed based on experiments with wideband noise spectra and thus may not be applicable for more natural and tonal sounds like those made by a saxophone (or any musical instrument for that matter).

Sounds perceived as rough are also described as harsh—ratings on the latter are typically found correlated with ratings on the former. However, acoustic analyses tend to associate harshness mainly with too much high-frequency energy (i.e., unpleasant). This is also evident in psycholinguistic studies of violin timbre (Fritz et al. 2012; Saitis et al. 2013) and voice quality (Garnier et al. 2007). Such descriptions include strident, shrill, piercing, harsh, and even nasal. Note that an implicit connection of roughness to energy in higher frequencies is also claimed by Helmholtz's hypothesis. Zacharakis et al. (2014, 2015) found that sounds with stronger high partials were described as rough or harsh and the opposite as rounded or soft and, to a lesser extent, as bright or sharp. They went on to suggest that spectral energy distribution is manifested primarily in descriptions of texture and not of brightness, which also relied on spectral detail. Rozé et al. (2017) showed that inappropriate bowing and posture coordination in cello performances resulted in energy transfer toward higher frequency harmonics, a decrease in attack time, and an increase in amplitude fluctuation of individual harmonics; this kind of timbre was perceived as harsh and shrill. Under optimal playing conditions, cello sounds were described as round.

A concept related to sensory dissonance, but distinct from roughness, is that of noisiness versus tonalness. The latter signifies the perception of strong stationary and near-periodic spectral components. As such, it has a close relation to pitch patterns. In this case, the timbre tends to be described as pure, clear or clean, and even bright. When random transients dominate the spectrum, the timbre tends to be described as noisy or blurry and messy. A dimension of tonal–noisy has been found for different types of timbres, including electroacoustic sounds (Sect. 5.3). However, specifically in bowed-string instruments, audible noise can still be present even when a clear and steady tonal component is established (Štěpánek 2006; Saitis et al. 2017). One source of such noise, sometimes described as rustle, is the self-excitation of subfundamental harmonics, particularly in the upper register (Štěpánek and Otcěnásek 1999). Another source is the differential slipping of bow hairs in contact with the string (McIntyre et al. 1981). In fact, adding such audible noise to synthesis models for instrumental sounds is known to enhance their perceived naturalness (Serra 1997).

Helmholtz (1877) and Lichte (1941) found that the predominance of odd harmonics in a spectrum (such as clarinet notes) elicits an impression of hollowness or thinness compared to sounds with more balanced spectral envelopes (such as bowed strings) that are perceived as full. Despite explicitly synthesizing odd and even harmonic spectra to test the thin–full hypothesis, von Bismarck (1974a) did not report any relation between those stimuli and his fullness factor. Hollowness has also been found connected to the amount of *phantom partials* (nonlinearly generated frequencies due to string tension modulation) in piano sounds (Bensa et al. 2005). A small number of phantom partials produces a hollow timbre; gradually increasing the presence of such partials gives a rounder timbre, but sounds with a very large number of phantom partials (i.e., more such partials in the upper register) can appear metallic and aggressive.

The mass dimension of Zacharakis et al. (2014) exhibited three strong correlations in the English listeners' group (results for the Greek group were less conclusive). Thickness and density increased with inharmonicity and with fluctuation of the spectral centroid over time and decreased with fundamental frequency. Similar to the first correlation, Bensa et al. (2005) observed that synthetic piano sounds with the least high-frequency inharmonic partials were perceived as poor, whereas increasing their number resulted in richer timbres. The second correlation appears to be in agreement with the connection between richness and high spectral variation reported for wind instruments by Kendall and Carterette (1993b) and for sustained instruments by Elliott et al. (2013) and may relate, at least partially, to multiple-source sounds with higher spectral flux values below 200 Hz that are perceived as fuller (Alluri and Toiviainen 2010).

The correlation between thickness/density and fundamental frequency found by Zacharakis et al. (2014) emerged largely due to the presentation of stimuli with different pitches. This acoustic interpretation of thickness/density alludes to an attribute of pure tones described by Stumpf (1890) as volume (*Tongröße* in German), which aligns inversely with pitch in that lower/higher pitches are larger/smaller. Together, the three attributes of volume, pitch, and loudness determine what Stumpf termed *tone color* (*Tonfarbe*). Rich (1916) provided empirical evidence that volume (he used the word extensity) can be distinct from pitch in pure tones. Terrace and Stevens (1962) showed that volume can also be perceived in more complex tonal stimuli, specifically, quarter-octave bands of pitched noise, and that it increases with loudness but decreases with pitch. Stevens (1934) observed that pure and complex tones further possess an attribute of density, which changes with loudness and pitch in a manner similar to perceptions of brightness: the brighter the tone, the louder and the less dense it is (Boring and Stevens 1936; cf., Zacharakis et al. 2014). Empirical observations of volume and density perceptions for pure tones have cast doubt on Schaeffer's (1966) claim that these have no timbre (Sect. 5.2.1).

Further experiments by Stevens et al. (1965) provided empirical support to Koechlin's claim that density is proportional to loudness and inversely proportional to volume (Sect. 5.2.1). An inverse relation between spectral centroid and volume was observed, which has been confirmed by Chiasson et al. (2017). They found that high energy concentrated in low frequencies tends to increase perceived volume, whereas low energy more spread out in higher frequencies tends to decrease it. Similarly, Saitis et al. (2015) showed that violin notes characterized as rich tended to have a low spectral centroid or stronger second, third, and fourth harmonics, or a predominant fundamental. Given that in harmonic sounds the fundamental is the lowest frequency, these findings generally agree with Helmholtz's (1877) claim that the stronger versus weaker the fundamental is relative to the upper partials, the richer versus poorer the sound is perceived.

5.4.2 Influence of Language and Culture

In the interlanguage study of Zacharakis et al. (2014, 2015), the overall configurational and dimensional similarity between semantic and perceptual spaces in both the English and Greek groups illustrates that the way timbre is conceptualized and communicated can indeed capture some aspects of the perceptual structure within a set of timbres, and that native language has very little effect on the perceptual and semantic processing involved, at least for the two languages tested. There also seems to be some agreement regarding the number and labeling of dimensions with studies in German (von Bismarck 1974a; Štěpánek 2006), Czech (Moravec and Štěpánek 2005; Štěpánek 2006), Swedish (Nykänen et al. 2009), and French (Faure 2000; Lavoie 2013). Chiasson et al. (2017) found no effect of native language (French versus English) on perceptions of timbral volume. All these studies were conducted with groups of Western listeners and with sounds from Western musical instruments. Further evidence of whether language (but also culture) influences timbre semantics comes from research involving non-Western listeners and non-Western timbres.

Giragama et al. (2003) asked native speakers of English, Japanese, Bengali (Bangladesh), and Sinhala (Sri Lanka) to provide dissimilarity and semantic ratings of six electroacoustic sounds (one processed guitar, six effects). Multidimensional analyses yielded a two-dimensional MDS space shared across the four groups and two semantic factors (sharp/clear and diffuse/weak) whose order and scores varied moderately between languages and related differently to the MDS space. For Bengali and Sinhala, both Indo-Aryan languages, the similarity between the respective semantic spaces was much stronger, and they correlated better with the MDS space than for any other language pair, including between the Indo-European English and Indo-Aryan relatives. Furthermore, the sharp/clear and diffuse/weak factors closely matched the semantic space of electroacoustic textures found by Grill (2012), whose study was conducted with native German speakers.

Alluri and Toiviainen (2010) found a three-dimensional semantic timbre space of activity (strong–weak, soft–hard), brightness (dark–bright, colorless–colorful), and fullness (empty–full) for Indian pop music excerpts rated by Western listeners who had low familiarity with the genre. Here timbre refers to timbral mixtures arising from multiple-source sounds. Both the number and nature of these dimensions are in good agreement with Zacharakis et al. (2014). Furthermore, similar semantic spaces were obtained across two groups of Indian and Western listeners and two sets of Indian and Western pop music excerpts (Alluri and Toiviainen 2012). Acoustic analyses also gave comparable results between the two cultural groups and between the two studies. Intrinsic dimensionality estimation revealed a higher number of semantic dimensions for music from one's own culture compared to a culture that one is less familiar with, suggesting an effect of enculturation. Furthermore, Iwamiya and Zhan (1997) found common dimensions of sharpness (sharp–dull, bright–dark, distinct–vague, soft–hard), cleanness (clear–muddy, fine–rough), and spaciousness (rich–poor, extended–narrow) for music excerpts rated separately by Japanese and Chinese native speakers (type of music used was not reported). These dimensions appear to modestly match those found by Alluri and Toiviainen (2010) and by Zacharakis et al. (2014).

Taken as a whole, these (limited) results suggest that conceptualization and communication of timbral nuances is largely language independent, but some culture-driven linguistic divergence can occur. As an example, Zacharakis et al. (2014) found that, whereas sharp loaded highest on the luminance factor in English, its

Greek equivalent *οξύς* (oxýs) loaded higher on the texture dimension of the respective semantic space. Greek listeners also associated *παχύς* (pakhús), the Greek equivalent of thick, with luminance rather than mass. Furthermore, a well-known discrepancy exists between German and English concerning the words *Schärfe* and sharpness, respectively (see Kendall and Carterette 1993a, p. 456). Whereas *Schärfe* refers to timbre, its English counterpart pertains to pitch. On the one hand, such differences between languages may not imply different mental (nonlinguistic) representations of timbre but rather reflect the complex nature of meaning.

On the other hand, there exists evidence that language and culture can play a causal role in shaping nonlinguistic representations of sensory percepts, for example, auditory pitch (Dolscheid et al. 2013). This raises a crucial question concerning the use of verbal attributes by timbre experts such as instrument musicians: To what extent does experience with language influence mental representations of timbre? Based on their findings, Zacharakis et al. (2015) hypothesized that "there may exist a substantial latent influence of timbre semantics on pairwise dissimilarity judgements" (p. 408). This seems to be supported from comparisons between general dissimilarity, brightness dissimilarity, and brightness scaling data by Saitis and Siedenburg (in preparation), but more research is needed to better understand the relationship between linguistic and nonlinguistic representations of timbre. Nevertheless, semantic attributes, such as brightness, roughness, and fullness, appear generally unable to capture the salient perceptual dimension of timbre responsible for discriminating between sustained and impulsive sounds (Zacharakis et al. 2015).

5.5 Timbre Semantics and Room Acoustics: Ambiguity in Figure-Ground Separation

Imagine yourself listening to recordings of the famous violinist Yehudi Menuhin (1916–1999) performing on his Khevenhüller Strad in different concert halls. Does your impression of the sound of the violin or the sound of Menuhin change from one recording or hall to another? The answer would be almost certainly yes. The perceived timbre of a sound is not only a result of the physical characteristics of its source: It is always influenced by the properties of the acoustic environment that connects the sound source and the listener. Putting it differently, in evaluating the timbre of a sound, listeners invariably evaluate timbral characteristics of the presentation space too. The influence of the latter on the spectral shape of a sound, as illustrated by the room acoustic transfer function (Weinzierl and Vorländer 2015), is manifested in a characteristic amplification or attenuation of certain frequencies, superimposed by an increasing attenuation of the spectral envelope toward higher frequencies due to air absorption. The extent of these effects can vary substantially from one space to another, depending on the geometry and materials of the room.

When listeners try to perceptually separate the properties of the sound source from the properties of the room, they face a situation that has been described as *figure-ground organization* in Gestalt psychology. Although its origins lie in visual scene analysis, organizing a perceptual stream into foreground (figure) and background (ground) elements has been shown to apply also in the auditory realm (Bregman 1990). Listeners can group foreground sounds across the spectral or temporal array and separate them from a background of concurrent sounds. When timbre acts as a contributor to sound source identity (Siedenburg, Saitis, and McAdams, Chap. 1), figure-ground segregation is generally unambiguous. A violin note will always be recognized as such, categorically as well as relative to concurrent notes from other instruments, regardless of the performance venue—excluding deliberate attempts to blend instrumental timbres (Lembke and McAdams 2015). However, figure-ground separation becomes more complicated when one looks beyond sound source recognition.

During language socialization of musicians or music listeners, where timbre functions as qualia (Siedenburg, Saitis, and McAdams, Chap. 1), there is not a single moment when a musical instrument is heard without a room acoustic contribution (except under anechoic room conditions). Even if the specific characteristics of the respective performance spaces are different, it can be assumed that common properties of *any* room acoustic environment (e.g., high-frequency spectral attenuation and prolongation by reverberation) will, to a certain degree, become part of the mental representation of an instrument's sound. It can be shown, for instance, that the early part of a room's reverberation tail tends to merge with the direct sound perceptually, increasing the perceived loudness of the sound rather than being attributed to the response of the room (Haas 1972). In addition, many musical instruments have their own decay phase, and with decay times of up to 3 s for violins on the open string (Meyer 2009), it becomes difficult to predict the extent to which listeners can successfully segregate the source and room streams when communicating timbral qualities.

The role of timbre in the characterization of room acoustic qualities has traditionally received little attention. In the current standard on room acoustic measurements of musical performance venues, there is not a single parameter dedicated to the timbral properties of the hall (ISO 3382-1:2009). However, recent studies have highlighted timbre as a central aspect of room acoustic qualities (Lokki et al. 2016), with brilliance, brightness, boominess, roughness, comb-filter-like coloration, warmth, and metallic tone color considered as the most important timbral attributes of a specific performance venue (Weinzierl et al. 2018a, b). The ways in which the semantics of spaces *interact* with the semantics of timbre and the extent to which figure-ground separation is reflected in the language of space and source are objects for future research.

5.6 Summary

Timbre is one of the most fundamental aspects of acoustic communication and yet it remains one of the most poorly understood. Despite being an intuitive concept, timbre covers a very complex set of auditory attributes that are not accounted for by frequency, intensity, duration, spatial location, and the acoustic environment (Siedenburg, Saitis, and McAdams, Chap. 1), and the description of timbre lacks a specific sensory vocabulary. Instead, sound qualities are conceptualized and communicated primarily through readily available sensory attributes from different modalities (e.g., bright, warm, sweet) but also through onomatopoeic attributes (e.g., ringing, buzzing, shrill) or through nonsensory attributes relating to abstract constructs (e.g., rich, complex, harsh). These metaphorical descriptions embody conceptual representations, allowing listeners to talk about subtle acoustic variations through other, more commonly shared corporeal experiences (Wallmark 2014): with reference to the human and nonhuman voice (*instruments are voices*), as a tangible object (*sound is material*), and in terms of friction (*noise is friction*). Semantic ratings and factor analysis techniques provide a powerful tool to empirically study the relation between timbre perception (psychophysical dimensions), its linguistic descriptions (conceptual-metaphorical dimensions), and their meaning (semantic dimensions).

Common semantic dimensions have been summarized as brightness/sharpness (or luminance), roughness/harshness (or texture), and fullness/richness (or mass) and correspond strongly, but not one-to-one, with the three psychophysical dimensions along which listeners are known to perceive timbre similarity. In some cases, the dimensions are relatively stable across different languages and cultures, although more systematic explorations would be necessary to establish a cross-cultural and language-invariant semantic framework for timbre. A recent study with cochlear implant listeners indicated a dimension of brightness and one of roughness in relation to variations in electrode position and/or pulse rate (Marozeau and Lamping, Chap. 10). Furthermore, notions of timbral extensity and density have been central to spectromorphological models of listening and sound organization (Sect. 5.2.1) and to theories of sound mass music (Douglas et al. 2017). More generally, timbre is implicated in size recognition across a range of natural (e.g., speech, animals; see Mathias and von Kriegstein, Chap. 7) and possibly even abstract sound sources (Chiasson et al. 2017).

Long-term familiarity with and knowledge about sound source categories influence the perception of timbre as manifested in dissimilarity ratings (McAdams, Chap. 2). An interesting question that has not been fully addressed yet is whether source categories further exert an effect on the semantic description of timbre, given the strong link between linguistic and perceptual representations. In this direction, Saitis and Siedenburg (in preparation) compared ratings of dissimilarity based on brightness with ratings of general dissimilarity and found that the former relied primarily on (continuously varying) acoustic properties. Could the mass dimension be more prone to categorical effects due to its connection with source size recognition? Closely related to this question is the need to specify the role of affective mediation in timbre semantics. For example, bright timbres tend to be associated with happiness, dull with sadness, sharp with anger, and soft with both fear and tenderness (Juslin and Laukka 2004). McAdams (Chap. 8) discusses the effect of timbral brightness on emotional valence in orchestration contexts.

Nonauditory sensory attributes of timbre exemplify a particular aspect of semantic processing in human cognition: People systematically make many crossmodal mappings between sensory experiences presented in different modalities (Simner et al. 2010) or within the same modality (Melara and Marks 1990). The notion of sound color, timbre's alter ego, is exemplified in terms such as the German *Klangfarbe* (*Klang* + *Farbe* = sound + color) and the Greek *ηχόχρωμα* [ichóchroma] (*ήχος* [íchos] + *χρώμα* [chróma] = sound + color) and is itself a crossmodal blend. In viewing timbre semantics through the lens of crossmodal correspondences, questions about the perceptual and neural basis of the former can thus be reconsidered. What timbral properties of sound evoke the analogous impression as touching a smooth surface or viewing a rounded form? Are perceptual attributes of different sensory experiences (e.g., a smooth surface and a rounded form) mapped to similar or distinct timbres? Are crossmodal attributes (e.g., smooth, rounded) a result of supramodal representations (Walsh 2013) or of direct communication between modalities (Wallmark 2014)? Addressing these questions requires a comprehensive examination of auditory-nonauditory correspondences, including the collection of behavioral and neuroimaging data from appropriate tasks that extend beyond the semantic differential paradigm.

Acknowledgements Charalampos Saitis wishes to thank the Alexander von Humboldt Foundation for support through a Humboldt Research Fellowship.

Compliance with Ethics Requirements Charalampos Saitis declares that he has no conflict of interest.

Stefan Weinzierl declares that he has no conflict of interest.

References

Abeles H (1979) Verbal timbre descriptors of isolated clarinet tones. Bull Council Res Music Educ 59:1–7

Albersheim G (1939) Zur Psychologie der Toneigenschaften (On the psychology of sound properties). Heltz, Strassburg

Alluri V, Toiviainen P (2010) Exploring perceptual and acoustical correlates of polyphonic timbre. Music Percept 27:223–242

Alluri V, Toiviainen P (2012) Effect of enculturation on the semantic and acoustic correlates of polyphonic timbre. Music Percept 29:297–310

Almeida A, Schubert E, Smith J, Wolfe J (2017) Brightness scaling of periodic tones. Atten Percept Psychophys 79(7):1892–1896

Bell R (2015) PAL: the percussive audio lexicon. An approach to describing the features of percussion instruments and the sounds they produce. Dissertation, Swinburne University of Technology

Bensa J, Dubois D, Kronland-Martinet R, Ystad S (2005) Perceptive and cognitive evaluation of a piano synthesis model. In: Wiil UK (ed) Computer music modelling and retrieval. 2nd international symposium, Esbjerg, May 2004. Springer, Heidelberg, pp 232–245

Bernays M, Traube C (2009) Expression of piano timbre: verbal description and gestural control. In: Castellengo M, Genevois H (eds) La musique et ses instruments (Music and its instruments). Delatour, Paris, pp 205–222

Bernays M, Traube C (2011) Verbal expression of piano timbre: multidimensional semantic space of adjectival descriptors. In: Williamon A, Edwards D, Bartel L (eds) Proceedings of the international symposium on performance science 2011. European Association of Conservatoires, Utrecht, pp 299–304

Bloothooft G, Plomp R (1988) The timbre of sung vowels. J Acoust Soc Am 84:847–860

Boring EG, Stevens SS (1936) The nature of tonal brightness. Proc Natl Acad Sci 22:514–521

Bowles EA (1954) Haut and bas: the grouping of musical instruments in the middle ages. Music Discip 8:115–140

Bowling DL, Purves D, Gill KZ (2018) Vocal similarity predicts the relative attraction of musical chords. Proc Natl Acad Sci 115:216–221

Bregman AS (1990) Auditory scene analysis. The perceptual organization of sound. MIT Press, Cambridge

Brent W (2010) Physical and perceptual aspects of percussive timbre. Dissertation, University of California

Burton H (2016) Menuhin: a life, revised edn. Faber & Faber, London

Chiasson F, Traube C, Lagarrigue C, McAdams S (2017) Koechlin's volume: perception of sound extensity among instrument timbres from different families. Music Sci 21:113–131

Cousineau M, Carcagno S, Demany L, Pressnitzer D (2014) What is a melody? On the relationship between pitch and brightness of timbre. Front Syst Neurosci 7:127

Daniel P, Weber R (1997) Psychoacoustical roughness: implementation of an optimized model. Acta Acust united Ac 83:113–123

Douglas C, Noble J, McAdams S (2017) Auditory scene analysis and the perception of sound mass in Ligeti's continuum. Music Percept 33:287–305

de Ceuster D (2016) The phenomenological space of timbre. Dissertation, Utrecht University

Disley AC, Howard DM, Hunt AD (2006) Timbral description of musical instruments. In: Baroni M, Addessi AR, Caterina R, Costa M (eds) Proceedings of the 9th international conference on music perception and cognition, Bologna, 2006

Dolan EI (2013) The orchestral revolution: Haydn and the technologies of timbre. Cambridge University Press, Cambridge

Dolscheid S, Shayan S, Majid A, Casasanto D (2013) The thickness of musical pitch: psychophysical evidence for linguistic relativity. Psychol Sci 24:613–621

Dubois D (2000) Categories as acts of meaning: the case of categories in olfaction and audition. Cogn Sci Q 1:35–68

Edwards RM (1978) The perception of trombones. J Sound Vib 58:407–424

Eitan Z, Rothschild I (2011) How music touches: musical parameters and listeners' audio-tactile metaphorical mappings. Psychol Music 39:449–467

Elliott TM, Hamilton LS, Theunissen FE (2013) Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. J Acoust Soc Am 133:389–404

Fastl H, Zwicker E (2007) Psychoacoustics: facts and models, 3rd edn. Springer, Heidelberg

Faure A (2000) Des sons aux mots, comment parle-t-on du timbre musical? (From sounds to words, how do we speak of musical timbre?). Dissertation, Ecoles des hautes etudes en sciences sociales

Fritz C, Blackwell AF, Cross I et al (2012) Exploring violin sound quality: investigating English timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties. J Acoust Soc Am 131:783–794

Gallese V, Lakoff G (2005) The brain's concepts: the role of the sensory-motor system in conceptual knowledge. Cogn Neuropsychol 22:455–479

Garnier M, Henrich N, Castellengo M et al (2007) Characterisation of voice quality in Western lyrical singing: from teachers' judgements to acoustic descriptions. J Interdiscipl Music Stud 1:62–91

Giragama CNW, Martens WL, Herath S et al (2003) Relating multilingual semantic scales to a common timbre space – part II. Paper presented at the 115th audio engineering society convention, New York, 10–13 October 2003

Grill T (2012) Perceptually informed organization of textural sounds. Dissertation, University of Music and Performing Arts Graz

Haas H (1972) The influence of a single echo on the audibility of speech. J Audio Eng Soc 20:146–159

Heise DR (1969) Some methodological issues in semantic differential research. Psychol Bull 72:406–422

Helmholtz H (1877) Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik, 4th edn. F. Vieweg und Sohn, Braunschweig. English edition: Helmholtz H (1954) On the sensations of tone as a physiological basis for the theory of music (trans: Ellis AJ), 2nd edn. Dover, New York

Isshiki N, Okamura H, Tanabe M, Morimoto M (1969) Differential diagnosis of hoarseness. Folia Phoniatr 21:9–19

Iwamiya S, Zhan M (1997) A comparison between Japanese and Chinese adjectives which express auditory impressions. J Acoust Soc Jpn 18:319–323

Jakobson R, Halle M (1971) Fundamentals of language, 2nd edn. Mouton, The Hague

Juslin PN, Laukka P (2004) Expression, perception, and induction of musical emotions: a review and a questionnaire study of everyday listening. J New Music Res 33:217–238

Kendall RA, Carterette EC (1993a) Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. Music Percept 10:445–468

Kendall RA, Carterette EC (1993b) Verbal attributes of simultaneous wind instrument timbres: II. Adjectives induced from Piston's orchestration. Music Percept 10:469–501

Kendall RA, Carterette EC, Hajda JM (1999) Perceptual and acoustical features of natural and synthetic orchestral instrument tones. Music Percept 16:327–363

Koelsch S (2011) Toward a neural basis of processing musical semantics. Phys Life Rev 8:89–105

Kursell J (2013) Experiments on tone color in music and acoustics: Helmholtz, Schoenberg, and Klangfarbenmelodie. Osiris 28:191–211

Lakoff G, Johnson M (2003) Metaphors we live by. University of Chicago Press, Chicago

Lavoie M (2013) Conceptualisation et communication des nuances de timbre à la guitare classique (Conceptualization and communication of classical guitar timbral nuances). Dissertation, Université de Montréal

Lemaitre G, Houix O, Misdariis N, Susini P (2010) Listener expertise and sound identification influence the categorization of environmental sounds. J Exp Psychol Appl 16:16–32

Lembke S-A, McAdams S (2015) The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds. Acta Acust United Ac 101:1039–1051

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cogn 21:1–36

Lichte WH (1941) Attributes of complex tones. J Exp Psychol 28:455–480

Lokki T, Pätynen J, Kuusinen A, Tervo S (2016) Concert hall acoustics: repertoire, listening position, and individual taste of the listeners influence the qualitative attributes and preferences. J Acoust Soc Am 140:551–562

Marozeau J, de Cheveigné A (2007) The effect of fundamental frequency on the brightness dimension of timbre. J Acoust Soc Am 121(1):383–387

McAdams S, Winsberg S, Donnadieu S et al (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res 58:177–192

McDermott JM, Lehr AJ, Oxenham AJ (2010) Individual differences reveal the basis of consonance. Curr Biol 20:1035–1041

McIntyre ME, Schumacher RT, Woodhouse J (1981) Aperiodicity in bowed-string motion. Acustica 49:13–32

Melara RD, Marks LE (1990) Interaction among auditory dimensions: timbre, pitch and loudness. Percept Psychophys 48:169–178

Meyer J (2009) Acoustics and the performance of music. Springer, Berlin

Miller R (1986) The structure of singing: system and art of vocal technique. Schirmer Books, New York

Moravec O, Štěpánek J (2003) Verbal description of musical sound timbre in Czech language. In: Bresin R (ed) Proceedings of the Stockholm Music Acoustics Conference 2003. KTH, Stockholm, p 643–646

Moravec O, Štěpánek J (2005) Relations among verbal attributes describing musical sound timbre in Czech language. In: Proceedings of Forum Acusticum Budapest 2005: the 4th European congress on acoustics. Hirzel, Stuttgart, p 1601–1606

Mores R (2011) Nasality in musical sounds – a few intermediate results. In: Schneider A, von Ruschkowski A (eds) Systematic musicology: empirical and theoretical studies. Peter Lang, Frankfurt am Main, pp 127–136

Nykänen A, Johansson Ö, Lundberg J, Berg J (2009) Modelling perceptual dimensions of saxophone sounds. Acta Acust United Ac 95:539–549

Osgood CE (1952) The nature and measurement of meaning. Psychol Bull 49:197–237

Porcello T (2004) Speaking of sound: language and the professionalization of sound-recording engineers. Soc Studies Sci 34:733–758

Pratt RL, Bowsher JM (1978) The subjective assessment of trombone quality. J Sound Vib 57:425–435

Pratt RL, Doak PE (1976) A subjective rating scale for timbre. J Sound Vib 45:317–328

Pressnitzer D, McAdams S (1999) Two phase effects in roughness perception. J Acoust Soc Am 105:2773–2782

Radocy RE, Boyle JD (eds) (2012) Psychological foundations of musical behavior, 5th edn. Thomas Books, Springfield

Reuter C (1997) Karl Erich Schumann's principles of timbre as a helpful tool in stream segregation research. In: Leman M (ed) Music, gestalt, and computing. Studies in cognitive and systematic musicology. Springer, Heidelberg, pp 362–372

Reybrouck M (2013) From sound to music: an evolutionary approach to musical semantics. Biosemiotics 6:585–606

Rich GJ (1916) A preliminary study of tonal volume. J Exp Psychol 1:13–22

Rioux V, Västfjäll D (2001) Analyses of verbal descriptions of the sound quality of a flue organ pipe. Music Sci 5:55–82

Rozé J, Aramaki M, Kronland-Martinet R, Ystad S (2017) Exploring the perceived harshness of cello sounds by morphing and synthesis techniques. J Acoust Soc Am 141:2121–2136

Saitis C, Fritz C, Guastavino C, Giordano BL, Scavone GP (2012) Investigating consistency in verbal descriptions of violin preference by experienced players. In: Cambouropoulos E, Tsougras C, Mavromatis P, Pastiadis K (eds) Proceedings of the 12th international conference on music perception and cognition and 8th triennial conference of the European Society for the Cognitive Sciences of Music, Thessaloniki

Saitis C, Fritz C, Guastavino C, Scavone GP (2013) Conceptualization of violin quality by experienced performers. In: Bresin R, Askenfelt A (eds) Proceedings of the Stockholm music acoustics conference 2013. Logos, Berlin, p 123–128

Saitis C, Fritz C, Scavone GP et al (2017) Perceptual evaluation of violins: a psycholinguistic analysis of preference verbal descriptions by experienced musicians. J Acoust Soc Am 141:2746–2757

Saitis C, Järveläinen H, Fritz C (2018) The role of haptic cues in musical instrument quality perception. In: Papetti S, Saitis C (eds) Musical Haptics. Springer, Cham, pp 73–93

Saitis C, Scavone GP, Fritz C, Giordano BL (2015) Effect of task constraints on the perceptual evaluation of violins. Acta Acust United Ac 101:382–393

Samoylenko E, McAdams S, Nosulenko V (1996) Systematic analysis of verbalizations produced in comparing musical timbres. Int J Psychol 31:255–278

Schaeffer P (1966) Traité des objets musicaux: essai interdisciplines. Editions du Seuil, Paris. English edition: Schaeffer P (2017) Treatise on musical objects: an essay across disciplines (trans: North C, Dack J). University of California Press, Oakland

Schneider A (1997) "Verschmelzung", tonal fusion, and consonance: carl stumpf revisited. In: Leman M (ed) Music, gestalt, and computing. Springer, Berlin, pp 117–143

Schumann KE (1929) Physik der Klangfarben (Physics of timbres). Habilitation, Universität Berlin

Serra X (1997) Musical sound modelling with sinusoids plus noise. In: Roads C, Pope S, Piccialli A, de Poli G (eds) Musical signal processing. Swets Zeitlinger, Lisse, pp 91–122

Simner J, Cuskley C, Kirby S (2010) What sound does that taste? Cross-modal mappings across gustation and audition. Perception 39:553–569

Slawson W (1985) Sound color. University of California Press, Berkeley

Smalley D (1997) Spectromorphology: explaining sound-shapes. Organised Sound 2:107–126

Solomon LN (1958) Semantic approach to the perception of complex sounds. J Acoust Soc Am 30:421–425

Štěpánek J (2006) Musical sound timbre: verbal description and dimensions. In: Proceedings of the 9th international conference on digital audio effects. McGill University, Montreal, p 121–126

Štěpánek J, Otčenásek Z (1999) Rustle as an attribute of timbre of stationary violin tones. Catgut Acoust Soc J (Series II) 3:32–38

Stevens SS (1934) Tonal density. J Exp Psychol 17:585–592

Stevens SS, Guirao M, Slawson AW (1965) Loudness, a product of volume times density. J Exp Psychol 69:503–510

Stumpf C (1890) Tonpsychologie (Psychology of sound), vol 2. Hirzel, Leipzig

Stumpf C (1898) Konsonanz und Dissonanz (Consonance and dissonance). Barth, Leipzig

Sundberg J (2013) Perception of singing. In: Deutsch D (ed) The psychology of music, 3rd edn. Academic, London, pp 69–105

Susini P, Lemaitre G, McAdams S (2012) Psychological measurement for sound description and evaluation. In: Berglund B, Rossi GB, Townsend JT, Pendrill LR (eds) Measurement with persons: theory, methods, and implementation areas. Psychology Press, New York, pp 227–253

Terrace HS, Stevens SS (1962) The quantification of tonal volume. Am J Psychol 75:596–604

Traube C (2004) An interdisciplinary study of the timbre of the classical guitar. Dissertation, McGill University

Thiering M (2015) Spatial semiotics and spatial mental models. Figure-ground asymmetries in language. De Gruyter Mouton, Berlin

Vassilakis PN, Kendall RA (2010) Psychoacoustic and cognitive aspects of auditory roughness: definitions, models, and applications. In: Rogowitz BE, Pappas TN (eds) Human vision and electronic imaging XV. SPIE/IS&T, Bellingham/Springfield, p 75270

von Bismarck G (1974a) Timbre of steady tones: a factorial investigation of its verbal attributes. Acustica 30:146–159

von Bismarck G (1974b) Sharpness as an attribute of the timbre of steady sounds. Acustica 30:159–172

Wake S, Asahi T (1998) Sound retrieval with intuitive verbal expressions. Paper presented at the 5th international conference on auditory display, University of Glasgow, 1–4 November 1998

Walker P (2016) Cross-sensory correspondences: a theoretical framework and their relevance to music. Psychomusicology 26:103–116

Wallmark Z (2014) Appraising timbre: embodiment and affect at the threshold of music and noise. Dissertation, University of California

Wallmark Z (2018) A corpus analysis of timbre semantics in orchestration treatises. Psychol Music. https://doi.org/10.1177/0305735618768102

Walsh V (2013) Magnitudes, metaphors, and modalities: a theory of magnitude revisited. In: Simner J, Hubbard E (eds) Oxford handbook of synesthesia. Oxford University Press, Oxford, pp 837–852

Webster J, Woodhead M, Carpenter A (1970) Perceptual constancy in complex sound identification. Br J Psychol 61:481–489

Weinzierl S, Lepa S, Ackermann D (2018a) A measuring instrument for the auditory perception of rooms: the Room Acoustical Quality Inventory (RAQI). J Acoust Soc Am 144:1245–1257

Weinzierl S, Lepa S, Schultz F et al (2018b) Sound power and timbre as cues for the dynamic strength of orchestral instruments. J Acoust Soc Am 144:1347–1355

Weinzierl S, Vorländer M (2015) Room acoustical parameters as predictors of room acoustical impression: what do we know and what would we like to know? Acoust Aust 43:41–48

Zacharakis A, Pastiadis K, Reiss JD (2014) An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. Music Percept 31:339–358

Zacharakis A, Pastiadis K, Reiss JD (2015) An interlanguage unification of musical timbre: bridging semantic, perceptual, and acoustic dimensions. Music Percept 32:394–412

Chapter 6
Neural Correlates of Timbre Processing

Vinoo Alluri and Sudarsana Reddy Kadiri

Abstract The brain is the most complex biological system that exists. Timbre, in its very nature, is a multidimensional concept with several levels of abstraction thus rendering the investigation of its processing in the brain extremely challenging. Timbre processing can be discussed in relation to levels of abstraction. Low- to mid-level representations can be associated with the neural representation of acoustic structure while high-level abstractions correspond to the neural representation of sound source properties. Furthermore, neural correlates of timbre can be broadly classified based on three stimulus categories, that is, those pertaining to music, speech, and environmental sounds. This chapter summarizes studies that have attempted to uncover neural correlates of varying levels of timbre abstractions. Finally, developments in methodological approaches are described, including the shift from univariate to multivariate statistical models, the employment of more naturalistic stimuli, and brain measurement paradigms from hitherto controlled auditory paradigms.

Keywords Acoustic feature decomposition · Dual-processing pathways · Hierarchical levels of abstraction · Multivoxel pattern analysis · Neuroimaging · Secondary auditory cortex

6.1 Introduction

Operationally, timbre can be defined as the attribute that discriminates sounds (such as speech, music, and environmental sounds) of equal pitch, loudness, and duration. In contrast to the latter facets of sound that rely on reasonably clearly defined

V. Alluri (✉) · S. R. Kadiri
International Institute of Information Technology, Gachibowli, Hyderabad, India
e-mail: vinoo.alluri@iiit.ac.in; sudarsanareddy.kadiri@research.iiit.ac.in

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_6

physical dimensions (e.g., pitch represented by periodicity and loudness represented by root-mean-squared energy), timbre is multidimensional, and to date there is a lack of consensus on the acoustic features that capture it exhaustively. Functionally, timbre is a key determinant of sound identity as it is the principal facet driving phonetic identity (such as vowels and consonants) in speech, musical instrument recognition (such as piano and violin), and identification of environmental sounds (such as car horns and closing doors). This distinction helps characterize timbre at varying levels of abstraction in the brain, which in turn stimulates hypotheses about how timbre is represented in the auditory system.

6.1.1 Levels of Abstraction

At low and middle levels of abstraction, the neural correlates of timbre have been associated with the representation of acoustic structure. On the other hand, the neural representation of sound source properties or sound sources (e.g., violin versus piano or human voices versus tool sounds) relates to high levels of abstraction. These high-level representations are characterized by perceptual constancy. For instance, a plucked sound of a violin string and a bowed sound of the same instrument are still perceptually categorized into one category, that is, a violin sound, despite the variance in articulation that results in varying timbral properties, which is also referred to as its “macrotimbre” (Sandell 1998). Similarly, a baby’s laughter and cry would still be categorized as human sounds albeit with varying emotional content. A higher-level or meta-level abstraction of timbre can be described in terms of timbral environments (Ferrer 2011), wherein an amalgamation of low-level to high-level representations of timbre can give rise to the ability to identify and categorize prototypical mixtures of sources. For example, humans are able to identify the overall emerging sound of a classical ensemble versus a big-band orchestra versus a hip-hop group; humans also categorize predominant environmental sounds emerging from a house by the sea versus those in the city versus those surrounding a busy restaurant kitchen.

An interesting analogy is drawn by Bregman (1990) to explain the phenomenon of emergence of a higher-order form from lower-order constituent elements. The author uses a knife as an example, wherein the molecules that constitute a knife are not sharp, however sharpness can be considered as an emergent property of a knife. Gjerdingen and Perrott (2008) describe global sound or overall timbre as an agglomerate “of spectral and rapid time-domain variability in an acoustic signal,” which is put together by the listener in a Gestalt-like manner that thereby enables listeners to identify, classify, and categorize, for example, the genre of the heard piece of music.

6.1.2 Hierarchical Processing

A question arises concerning the direction of hierarchical processing in the auditory brain, that is, if properties of sound are perceived in a bottom-up or top-down fashion. One could surmise that sounds that are novel and unnatural would be processed in a bottom-up fashion as a mental representation would not yet exist. On the other hand, sounds that are either familiar or more relevant (e.g., human voices versus synthesized timbres) would engender top-down processing by selective attention or weighted attention on the most relevant acoustic features. This dichotomy does not imply that a listener employs one of the two approaches for sensory perception but that perceptual mechanisms develop during the lifespan based on learning, familiarity, and conditioning of the mind, or by interaction with the environment. As such, the direction of hierarchical processing is chosen subconsciously based on the context and state of the listener. This view is analogous to the Giordano et al. (2010) description of young perceivers (or children) and adults wherein young perceivers try to avoid perceptual errors (thereby employing a bottom-up approach) by focusing on sensory content, whereas adults automatically employ a top-down approach to make the most educated guess even in the absence of accurate information (analogous to Helmholtz's theory of unconscious inference).

These distinctions can be discussed in terms of hierarchical feature-processing models of the auditory brain (Leaver and Rauschecker 2010). The ideology behind this putative hierarchy stems from research on the visual system, wherein information is broken down into its basic elements and is subsequently integrated into a complex representation of the image in the brain (Wessinger et al. 2001). Similarly, in the auditory modality, low-levels and mid-levels of abstraction capture acoustic structure and perceivable dimensions, respectively, which in turn contribute to higher levels of abstraction that enable the listener to make complex timbral judgements (qualifying, classifying, or categorizing the sound source). One of the earliest functional magnetic resonance imaging (fMRI) studies that investigated neural pathways responsible for processing complex sounds demonstrated the presence of two distinct pathways (Rauschecker 1998) (see Fig. 6.1). Both pathways stem from the core auditory areas (i.e., the primary auditory cortex A1 and rostral area) but fork out ventrally and dorsally only to converge later in the prefrontal cortex. This hierarchically organized object-processing pathway is associated with the anteroventral auditory pathway, which is key in processing auditory objects (what is the sound?) in contrast to the dorsal auditory pathway that is responsible for processing spatial aspects (where is the sound?) (Rauschecker and Scott 2009; Bizley and Cohen 2013).

Specifically, the ventral pathway branches out from A1 to the lateral belt of the superior temporal cortex while the dorsal pathway connects to the posterior parietal areas. The dorsal stream is purported to process spatial information related to sounds. The ventral stream plays a significant role in processing auditory patterns of complex sounds and, hence, is key in processing timbre. Furthermore, evidence

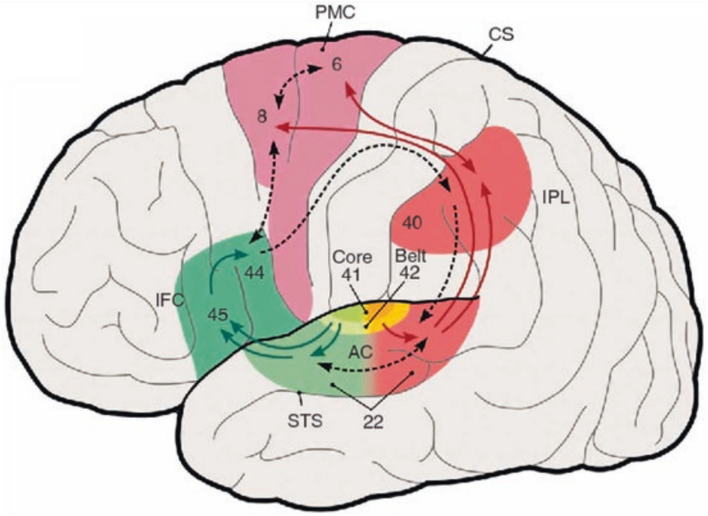

Fig. 6.1 Dual auditory processing scheme of the human brain. Anteroventral (*green*) and posterodorsal *(red*) streams originating from the auditory belt. *AC*, auditory cortex; *CS*, central sulcus; *IFC*, inferior frontal cortex; *IPL*, inferior parietal lobule; *PMC*, premotor cortex; *STS*, superior temporal sulcus. *Numbers* correspond to Brodmann areas. (Reproduced from Rauschecker and Scott 2009; used with permission of the publisher, Nature Neuroscience)

suggests that processing of levels of abstraction from low to high varies as a function of distance from the primary auditory cortex in the anterolateral direction (see Sect. 6.2.1 for more details).

6.1.3 Types of Stimuli

Historically, several neuropsychological studies on timbre perception (summarized in Sect. 6.2) have relied on the observation of neural encoding of carefully synthesized sounds and changes in them in order to unravel low-level processing of timbre in the brain. This reductionist approach allows one to pinpoint neural responses to systematic manipulations of particular aspects of the sound. However, this approach renders the relevance of the results questionable since the vast majority of the human species almost never hears these synthetic sounds or natural sounds in isolation, thereby decreasing the behavioral relevance of the neural findings. Santoro et al. (2014) argue that our brains have evolved to efficiently encode sounds in natural settings. Hence, setting an appropriate, feasible auditory context is crucial in understanding how the brain processes any stimulus. On the other hand, studies that have investigated categorical encoding of timbre (e.g., representation of vocal sounds versus musical instruments) with real-world sounds give us more information on how the auditory system has evolved to efficiently process the sounds that are of most importance to a species.

Neuroscientific research on timbre can be broadly classified into three categories: neural correlates of timbre in the context of music, speech, and environmental sounds. Low levels of abstraction of timbre in the context of speech have been associated with phonetic identity while high levels of abstraction have been representative of speaker identity (Town and Bizley 2013). In music, timbre studies have focused on either low-level to mid-level features associated with the perceived similarity in acoustic structure within (or across) instrument categories or on high-level abstraction in terms of perceptual constancy or invariance as described above. Environmental sounds can be characterized as an amalgamation of units of sounds that do not particularly change over time and are relatively static, leading to mid-level representations that reflect more complex acoustic regularities (Mlynarski and McDermott 2018).

6.1.4 Measurement Modalities and Modeling Approaches of Brain Activity

The majority of neuroscientific studies have employed fMRI as a means to observe brain activity (note that all abbreviations appear in Table 6.1). Responses measured using fMRI serve as an indirect measure of brain activity and as a result offer a window into slow-moving changes in the brain (around 3 s) and fMRI helps to address *where* in the brain changes occur due to stimulus changes.

Table 6.1 Abbreviations

DMN	Default mode network
EEG	Electroencephalography
fMRI	Functional magnetic resonance imaging
HG	Heschl's gyrus
IC	Inferior colliculus
MEG	Magnetoencephalography
MMN	Mismatch negativity
MTG	Middle temporal gyrus
MVPA	Multivariate pattern analysis
PET	Positron emission tomography
PT	Planum temporale
pMFG	Posterior middle frontal gyrus
pMTG	Posterior middle temporal gyrus
SOC	Superior olivary complex
STC	Superior temporal cortex
STG	Superior temporal gyrus
STRF	Spectrotemporal receptive field
STS	Superior temporal sulcus

Positron emission tomography (PET) serves as another means of localizing brain responses with greater spatial resolution than temporal resolution, but it is uncommon in human studies due to the need to inject radioactive tracers into participants to track metabolic activity. On the other hand, electroencephalography (EEG) is a direct measure of electrical brain activity represented by changes in scalp voltage. However, this method suffers from poor spatial resolution as the activity is picked up only from the scalp and is unable to capture neuronal activations of deep brain structures. Hence, EEG offers a window into understanding *when* the brain responds to changes. Magnetoencephalography (MEG) is the magnetic counterpart of EEG with greater potential for neuronal source localization.

While the majority of neuroimaging studies have employed univariate methods, recently there has been an increase in multivariate statistical approaches, namely, *multivariate pattern analysis* (MVPA). This method allows identification of distributed patterns that are representative of stimulus-evoked activations (Kriegeskorte et al. 2006; Stelzer et al. 2013) and the term is synonymous with *multivoxel pattern analysis*. This approach lies more in the domain-integrative approach of brain functioning, which aims to identify interactions between regions that allow integrated functioning (rather than the traditional segregation-based approach, which aims to identify local regions that are specialized for a particular task). MVPA has been further advocated as more sensitive than typical univariate analyses (Davis et al. 2014; Allen et al. 2017).

Univariate models are representative of a restricted kind of encoding model, that is, models that enable prediction of brain activity using stimulus properties typically at a voxel-level. Decoding approaches, on the other hand, allow prediction of stimulus features or properties based on brain activity (see Fig. 6.2) and are fast gaining popularity. MVPA is typically used in decoding approaches. Finally, the combination of encoding and decoding approaches permits capitalization of their relative

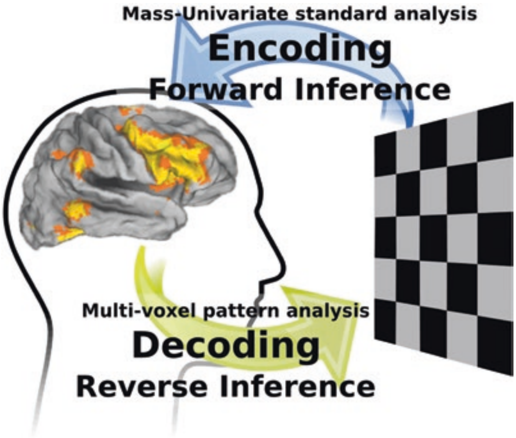
Mass-Univariate standard analysis
Encoding
Forward Inference
Multi-voxel pattern analysis
Decoding
Reverse Inference

Fig. 6.2 Schematic depiction of the distinction between encoding and decoding in brain imaging. The figure is for representational purposes only; colors indicate general activity and do not correlate with only encoding or only decoding activities. (Reproduced from Varoquaux and Thirion 2014; used with permission of the publisher, GigaScience)

strengths by providing a sanity check of the results, which boosts the robustness of the results and hence the conclusions drawn from them.

The following sections are organized to address neural processing of timbre in light of the aforementioned levels of abstraction and methods. Section 6.2 describes neural correlates of timbre related to lower levels of abstraction. The subsequent Sections 6.3 and 6.4 deal with high-level neural representations of timbre ranging from sound source identification to encoding global sound qualities.

6.2 Representation of Acoustic Structure

Transduction of raw auditory signals is a dynamic and rapid process, starting in the cochlea and leading to more compact and abstract representations as it moves along the primary auditory pathway (Mlynarski and McDermott 2017). Low-level representations of auditory signals have been typically characterized by using a variety of feature-decomposition models. These models are developed based on identifying optimal temporal and spectral resolutions that best fit the observed brain responses and are further validated with new, independent brain data (Santoro et al. 2014).

A group of decompositions has traditionally dominated computational models of the auditory system and comprises the spectrotemporal receptive field (STRF) (Miller et al. 2002) and modulation spectrum or related representations (Santoro et al. 2014) (see Elhilali, Chap. 12). These features can be termed *low level* to emphasize their relevance to the characterization of early processing in the cortical auditory stream and to distinguish them from source-related high-level features that are assumed to be object representations. However, such low-level STRF representations have been found to best capture neuronal encoding in the brain stem, but they fail to completely capture encoding in the auditory cortex as neurons there are found to be sensitive simultaneously to multiple stimulus features, thereby giving rise to mid-level representations that respond to combinations of low-level features. Hence, mid-level representations are akin to compact summaries by means of an efficient recoding of low-level features. This process can be construed as an intermediate stage between peripheral processing and perceptual decision-making stages (McDermott and Simoncelli 2011). For example, McDermott et al. (2013) describe sound *texture* as mid-level representation, which is based on time-averaged summary statistics of the auditory signal.

6.2.1 Low-Level Timbral Representations

Low-level features can be thought of as those that are perceived in a bottom-up fashion without a need for domain-specific knowledge or context. The cochlea performs frequency decomposition of the sound by applying varying shapes of filters based on frequency: that is, narrowband frequency filters for low frequencies and wideband frequency filters for higher frequencies. The information is then relayed by the auditory nerve fibers via the superior olivary complex (SOC) to brainstem structures, the inferior colliculi (IC). Representation of acoustic structure in terms of coarse spectral decompositions already starts in the IC before reaching the primary auditory cortices via the thalamus (Rodriguez et al. 2010; Ress and Chandrasekaran 2013). Particularly, the IC was found to have finer encoding of low and high frequencies than intermediate ones. On the other hand, the primary auditory cortex has a more distributed and pronounced encoding of frequency from high to low to high running along the posterior-anterior axis in a V-shaped pattern surrounding Heschl's gyrus (Formisano et al. 2003; Humphries et al. 2010). Frequency encoding along the primary auditory pathway can be understood as spectral representations of incoming sound at various resolutions that result in low-level to mid-level representations. The distribution of neuronal activity across this tonotopic map later gives rise to various timbral properties related to spectral shape.

Animal models have revealed widespread neural sensitivity to sound properties across the auditory cortex (Town and Bizley 2013). Apart from the well-established principle of tonotopic representation (the topographical representation of frequency), temporal information of the sounds is extracted by amplitude modulation filter banks that are already present in the inferior colliculus, as found in rodent, cat, and primate midbrains (Langer 2009; Baumann et al. 2011). Specifically, there exists evidence for an orthogonal representation of amplitude modulation rate and frequency. In primates, the topographic organization of amplitude modulation rate is at right angles to the already established frequency gradient (Baumann et al. 2015). Particularly, regions posterior to the medial Heschl's gyrus were associated with a preference for fast temporal rates, whereas regions anterolateral to the medial Heschl's gyrus showed a preference for low temporal rates. Overall, the patterns that emerged were found to be arranged in concentric isorate bands that are mirror symmetric across both hemispheres. These observations are consistent with those in the human auditory cortex with preference for the higher rates in the posteromedial cortex surrounded by anterior and lateral areas demonstrating a preference for lower temporal rates. The distribution of neuronal activity across this amplitude modulation rate map might later give rise to timbral properties based on temporal stimulus features.

Lesion studies have typically demonstrated the importance of the right temporal lobe in timbre processing. However, lesion studies also report deficits in timbre processing due to left hemispheric lesions and show that the nonprimary auditory cortex plays a significant role (Samson et al. 2002; Samson 2003). There is evidence supporting the view that the auditory cortex decomposes the incoming sound signal in parallel at multiple temporal and spectral resolutions (Samson et al. 2011; Santoro et al. 2014) with a left-hemispheric bias for finer temporal resolution and right-hemispheric bias for finer spectral resolution (Zatorre and Belin 2001). The anterior auditory regions, in particular, have been found repeatedly to contribute to the finer encoding of spectral content, which comprises key timbre features that are then integrated at a later stage for higher-order sound-source representations.

Despite Leaver and Rauschecker's (2010) corroboration of the notion of a hierarchically organized pathway for the processing of complex sounds along the anteroventral auditory cortex, they also reported that the posterior auditory cortex (i.e., closer to auditory core cortex) is associated with processing low-level acoustic features that are not specific to any category of sounds, thereby placing it lower on the hierarchical auditory processing pathway.

Research on the neural correlates of timbre in music has revolved around instrument categorization or reactions to controlled alterations of instrument-like sounds. Since the former relates to source identification, the latter will be discussed in this section. Overall, passive listening to synthesized timbres that differ in carefully manipulated acoustic features has been associated with bilateral activations in auditory areas (Menon et al. 2002; Caclin et al. 2008). Electrophysiological studies that employed the mismatch negativity (MMN) paradigm to investigate the neural correlates of perception of timbre have hinted at a high correlations between the amplitude of brain responses and the magnitude of changes in certain timbral features. A discernable change in a sequence of sounds elicits an MMN and often reflects preattentive, automatic processes. These changes correspond to alterations in features, such as brightness (Toiviainen et al. 1998), that are represented by the spectral center of mass in monophonic sounds and by the attenuation of even harmonics (Caclin et al. 2006; Caclin et al. 2008) of carefully synthesized sounds.

The spectral centroid or spectral center of mass is a summary feature of the spectrum (Siedenburg, Saitis, and McAdams, Chap. 1) that has frequently been equated to perceptual brightness (see McAdams, Chap. 2; Saitis and Weinzierl, Chap. 5). It has often been referred to as a low-level feature despite it being based on a summary statistic, which could be termed a mid-level representation of acoustic structure. This example also highlights that it is hard to draw a clear distinction between low-level and mid-level features. In fact, timbre features, such as the spectral centroid and the attack time, may represent mid-level perceptual attributes, such as brightness and percussiveness, which in turn aid in categorizing and recognizing the sound source (albeit typically only in the context of isolated musical-instrument sounds or synthesized sounds).

Caclin et al. (2006, 2008) employed the Garner interference approach that helps assess the separability of stimulus dimensions. In other words, it allows one to assess if processing changes in one auditory dimension (e.g., spectral center of gravity) are affected by change in an unattended auditory dimension (e.g., attack time). They found that three major dimensions of timbre, namely, attack time, spectral centroid, and spectrum fine structure, are distinctly processed at very initial stages but interact at a later stage (delay of the order of 250 ms).

Along similar lines, Allen et al. (2017) attempted to identify brain regions specific to the processing of variations in either pitch (fundamental frequency) or timbre (spectral centroid). In contrast to previous studies that supported modular processing of these variations, they observed a large degree of spatial overlap, especially in Heschl's gyrus (HG), thereby corroborating the presence of shared neuronal substrates in encoding pitch and timbre. However, MVPA, which is more sensitive than univariate approaches, revealed distinct combinations of neuronal populations within these overlapping regions that are associated with either pitch or timbre processing. In other words, pitch and timbre as perceptual constructs are not independent of each other. This has been reflected in previous psychoacoustic studies that report variations in the pitch of harmonic complex tones affecting the perception of timbre and vice-versa (Allen and Oxenham 2014).

In addition, studies that have investigated the perceptual and acoustic correlates of timbre of global sound found that pitch content is a crucial component in judgements concerning the perceived brightness of short (1 s) musical excerpts (Alluri and Toiviainen 2010). This result suggested that pitch processing appears to be embedded in timbre processing, which as a result would recruit primary auditory areas in both the anterior and posterior directions due to the aforementioned V-shaped frequency encoding that surrounds HG. In addition, the presence of high notes in the stimuli may render them perceptually brighter. In contrast, some stimuli that contained rapidly repeating percussive sounds lacking a prominent pitch in the higher registers were rated as less perceptually bright despite an increase in the energy contained in the higher end of the spectrum due to those sounds (Alluri and Toiviainen 2010).

6.2.2 Mid-Level Timbral Representations

Paralleling developments in the understanding of the cortical processing of visual objects, McDermott and Simoncelli (2011) proposed that an intermediate level of representation in the auditory system relies on statistics of incoming low-level decompositions. This intermediate level of representation transforms acoustic structure into knowledge of the environment—the sound source. *Sound textures* are distinguished by the collective temporal homogeneity of acoustic events. For example, naturally occurring environmental sounds, such as rainstorms and galloping horses, can be characterized by similitude in their acoustic properties over time, thereby allowing us to categorize and identify the acoustic event and respective sources. Sound texture categorization may be better characterized by long-term statistical properties rather than fine spectrotemporal modulations as observed in isolated sounds (Theunissen and Elie 2014). The middle superior temporal sulcus (STS) is purported to play an intermediary role by integrating spectral shape and acting as an anatomical and computational link relaying information from primary auditory cortex (processing basic source attributes such as pitch height) to anterior and inferior temporal lobe regions that are involved in source recognition (Warren et al. 2005). The STS relays information processed in the primary auditory cortex to the lateral superior temporal plane areas and to inferior and anterior temporal regions for the purpose of categorization (Belin et al. 2000; Warren et al. 2005).

Neuroimaging studies (fMRI and PET) have helped to localize regions of the brain that deal with timbre processing of isolated sounds that were synthesized by systematically controlling low-level to mid-level features. Menon et al. (2002) investigated brain responses to melodies with controlled, synthesized timbres varying in their fundamental acoustic dimensions, namely, attack time, spectral centroid, and spectral flux, which were derived from previous psychophysical timbre experiments (McAdams et al. 1995). Timbre processing was associated with activations in the mid-sections of the superior temporal gyrus (STG), STS, and adjoining insular cortex with greater activation for the spectrally more complex sounds. They further demonstrated the asymmetry in hemispheric activations with more posterior left hemispheric activation versus more anterior activations in the right hemispheric temporal lobes. On the other hand, selectively attending to timbre (melody played by two synthesized oboe sounds—one possessing more harmonics than the other) while pitch and rhythmic content varied simultaneously was associated with activations in the right superior and middle frontal gyri (Platel et al. 1997). Platel et al. (1997) hypothesize that this likely reflects a directing of attention to one source or the other rather than to the actual encoding of timbre, since these regions are part of the attention-processing network (Petersen and Posner 2012).

In sum, a majority of studies suggest the ventral pathway as the main pathway that encodes timbre at various processing stages. Low-level timbre processing already begins in the primary auditory pathway by means of spectral decompositions followed by distributed representation in the auditory cortices. Stemming from the primary auditory cortex, mid-level representations are formed in the secondary areas such as STS and anterior portions of the STG with a right-hemispheric bias. Mid-level representations can be considered analogous to descriptive summaries (such as brightness, roughness) that constitute perceivable dimensions of auditory sensation and subsequently contribute to high-level representations of sound sources.

6.3 Representation of Sound Sources

Listeners effortlessly differentiate between sources such as human voices and tool sounds. These categorical distinctions constitute one of the most general attributes of a sound source and have been the object of neuroimaging research for almost two decades. The brain's ability to modularize the incoming auditory stream into perceptual units or "auditory objects" based on timbral properties is fundamental in parsing and decoding this stream. Interpreting information as a result of integrating these perceptual units provides information about where, from whom or what it came (source identity), in addition to its content (from phonemes to words to emotional connotations). For example, categorical distinctions can be related to instrument recognition in music, or vocal tract length (short versus long), phoneme, gender, and emotions in speech, and material, hollowness, size, and shape of vibrating objects in environmental sounds.

As Bizley and Cohen (2013) rightly point out, there is no absolute definition of what these auditory objects are. However, the general accepted notions relate to perceptual constructs of sound that can be associated with a specific source (e.g., a honk associated with a car). The formation of these objects in the brain is also the result of attentional processes with weighting that leads the listener to focus on certain aspects in order to efficiently encode them. This process is akin to Goldstone's (1998) description of categorical perception as a result of attentional weighting. For instance, Bizley and Cohen (2013) give an example of how, while listening to an orchestra, we can selectively either attend to or ignore sections of it (such as the first or second violin sections). This process might result in the recruiting of attention networks in addition to timbre-processing regions. For example, Deike et al. (2004) observed (via fMRI) greater involvement of the left auditory cortex than the right in processing sequences with alternating timbres (complex sounds with an organ-like spectral envelope alternating with a trumpet-like spectral envelope) compared to the condition wherein the same sound was presented repeatedly. The authors interpreted this as evidence for a selective involvement of posterior regions of the left primary auditory cortex for auditory stream segregation based on timbre cues (spectral differences in this case). Since existing evidence points to a left-hemispheric bias for processing temporal sequences, this result is not surprising considering that the task at hand required focus on the temporal sequence for target sound detection.

6.3.1 Categorization of Musical Instruments

A majority of lesion studies provide support for right-lateralized bias mostly in the anterior regions in the categorical processing of musical instruments (Samson 2003). In line with the putative finer spectral information processing capabilities found in the right temporal lobe, EEG studies have reported larger recorded electrical activity over the right than the left hemisphere (Crummer et al. 1994; Auzou et al. 1995). Evidence from fMRI studies indicates that right superior temporal regions are active in the processing of musical instruments versus other categories (such as speech, animal vocalizations) (Leaver and Rauschecker 2010). In a MEG study, musical training was a modulatory factor in processing timbre changes and led to enhanced cortical representations for the timbre of a listener's own musical instrument (Pantev et al. 2001). Crummer et al. (1994) concluded from their EEG study that listeners with musical experience and training were better able to discriminate timbre changes, which was demonstrated by quicker reaction times as represented by shorter latency in the EEG responses.

Leaver and Rauschecker (2010) observed encoding of relevant sounds (human speech and musical instruments) versus nonrelevant sounds (songbird or animal vocalizations) in the anterior superior temporal cortex (STC). Furthermore, in line with previous studies, the right anterior superior temporal plane was found to preferentially encode musical instrument timbre. Additionally, the left STC was found to selectively process acoustic-phonetic content of human speech; the anterior regions were proposed to encode whole words in line with previous findings related to a left-hemispheric bias in speech processing. No regions were found to be selectively responsive to nonrelevant sounds, thereby supporting the notion that the human auditory cortices have indeed evolved to efficiently encode relevant stimuli with dedicated neuronal populations.

Along similar lines, Staeren et al. (2009) used acoustically matched stimuli (such as guitars, cats, and singers) to explore category representation in the auditory cortex. As sounds across acoustic categories were matched in pitch, timbre was the key perceptual determinant of sound category. It was found that spatial patterns of activation differentiated the three acoustic categories in higher auditory areas, including antero-lateral Heschl's gyrus (HG), the planum temporale (PT), and the posterior STG and/or STS.

6.3.2 Categorization of Human Speech

There exists converging evidence of right-hemispheric auditory region dominance in the processing of holistic aspects of sound from vowel classes in speech (Obleser et al. 2006) to speaker identification (Formisano et al. 2008). Specifically, the anterior STG has been found to play a key role in global processing of sound properties. A widely distributed network of regions, including the anterior HG, PT, STS, and STG, participate in the identification of phonetic content (e.g., what is being said) or categorizing speaker properties. More details concerning timbre in speech can be found in Mathias and von Kriegstein (Chap. 7).

From the aforementioned view of hierarchical processing of auditory information, there is consensus that the ventral auditory pathway is key in processing auditory objects due to its involvement in processing the nonspatial aspects of the sound (what is the sound?) versus spatial aspects (where is the sound?), which are processed in the dorsal auditory pathway (Bizley and Cohen 2013). There is evidence, however, of the encoding of spatial information in the ventral pathway and of auditory objects in the dorsal pathway (Bizley et al. 2009; Rauschecker 2012), thereby suggesting an interplay of spatial and nonspatial information during the process of creating consistent perceptual representations. Moving along the pathway can be thought of as moving along an anterior gradient representing levels of abstraction starting from low levels in A1 to high levels of categorical representations in STG and projecting onto the prefrontal cortex or, alternatively, projecting in both anterior and posterior directions (Giordano et al. 2013; Giordano et al. 2014).

In an fMRI study, Kumar et al. (2007) examined the flow of information using dynamic causal modeling and noted that timbre information originates in Heschl's gyrus and flows in a serial fashion to the planum temporale and then to STG. The stimuli used included either harmonic or noise-like synthesized sounds with various superimposed spectral envelopes. They concluded that spectral envelope encoding was already performed by the time information reaches the planum temporale from HG. However, this result has to be tested in the context of natural sounds.

6.3.3 Categorization of Living, Nonliving, and Human-Action Sounds

Giordano et al. (2013) demonstrated the encoding of auditory objects ranging from living sounds (e.g., clapping hands) to nonliving sounds (e.g., crumpling paper bag), human-action sounds (e.g., hammering nail), vocal animal and human sounds, and nonvocal sounds using MVPA in an fMRI study (Giordano et al. 2013). While low-level features were encoded in wide areas of the auditory cortex, abstract categories of living sounds and human sounds were represented in the posterior areas, including the planum temporale. This result suggests that the direction of information abstraction extends both in anterior and posterior directions starting from A1. Another fMRI study on musical imagery, which dealt with similarity judgements of either played or imagined sounds, found a slight right-hemispheric predominance in addition to posterior temporal areas, thereby lending further support to these bidirectional processing pathways (Halpern et al. 2004).

In a subsequent study, Giordano et al. (2014) found the left posterior middle frontal gyrus (pMFG) to be key in processing sound source identity in both passive and active listening conditions unaffected by the sound category (environmental sounds, music-instrument tones, human vocalizations). This unexpected finding was explained in light of the role of pMFG in general executive-control function operations and short-term working memory processes. The presentation of isolated sounds belonging to various categories for the purpose of determining source-identity information would indeed engage such regions and would rely on low-level sound structures due to the absence of contextual information.

A novel hypothesis-free study on identifying selective neuronal responses to natural sounds, including music, speech, and environmental sounds revealed neuronal selectivity throughout the auditory cortex (Norman-Haignere et al. 2015). The authors employed an elegant data-driven approach to identify distinct neuronal populations in the auditory cortex that were responsible for encoding a variety of naturally occurring sounds. As a result, in addition to reinforcing the previously suggested V-shaped tonotopic gradient in the Heschl's gyrus (Humphries et al. 2010), they observed that the anterior regions of the auditory cortex were music-selective, and selective speech encoding was associated with lateral regions.

In light of the dual-stream hypothesis, several studies on human sound identification demonstrate similarities in regions that process human actions versus nonliving sources. While the former categories recruit an *action-sound network* with regions belonging to the dorsal stream, the processing of sounds from the latter categories relies on the ventral pathway (Lemaitre et al. 2017). Lemaitre and colleagues also suggest that environmental sound processing is coupled with the process of identification of underlying processes. Environmental sounds can further be divided into living (e.g., animal sounds) versus nonliving sounds (e.g., tool sounds). Animal vocalizations activate the middle regions of bilateral STG more than nonliving sounds, whereas the left posterior middle temporal gyrus (MTG) and frontoparietal regions were more active in the reverse condition (Lewis 2005). Interestingly, several studies have reported selective encoding of tool sounds in posterior MTG (Beauchamp et al. 2004; Lewis et al. 2006). Doehrmann et al. (2008) reported that the position of the posterior middle temporal gyrus (pMTG) in the hierarchical processing pathway might indeed be higher than that of the STG. They further contend that the pMTG selectively processed tool sounds in addition to left-hemispheric regions while the anterior STG selectively adapted to animal sounds. These results lend further support to the evolutionary adaptation of human auditory cortex for living sounds.

A PET study revealed that passive listening to scrambled environmental sounds (e.g., sewing sounds, cat sounds) versus their unscrambled counterparts elicited activations in right anterior STG/STS and inferior frontal gyrus (Engelien et al. 2006). The scrambled sounds were considered meaningless and unnatural, which would therefore elicit bottom-up processing. On the other hand, the unscrambled sounds elicited greater responses in anterior left-hemispheric auditory regions.

Overall, the level of abstraction in auditory processing increases as a function of distance from primary auditory cortices (Lemaitre et al. 2017). The anterior STG appears to be vital in the representation of relevant sound sources. Furthermore, in certain instances, source identity processing involves frontal cortical regions in addition to secondary auditory regions. The recruitment of frontal regions is attributed to working memory and attentional networks due to the nature of experimental settings (such as category discrimination tasks).

So far we have summarized several neuroimaging studies that have carefully attempted to disentangle the neural correlates of timbre at various levels of abstraction. The experimental paradigms typically utilized in neuroimaging studies rely on highly reduced auditory stimuli and controlled experimental designs. Such paradigms fail to emulate real-world situations wherein the auditory system is constantly bombarded with continuous streams of sensory information; hence, these paradigms reveal only an incomplete picture of the neural mechanisms involved in the processing of realistic stimuli. In the visual modality, recent evidence suggests that the brain processes visual stimuli presented in a more ecological setting differently than when presented in conventional controlled settings (Hasson et al. 2004). Assuming that this finding is generalizable across sensory modalities, one could expect that the majority of studies in the auditory modality, in which acoustic features of sounds were artificially manipulated or were presented in isolation, may have revealed an incomplete picture of neural correlates of timbre. Hence, setting an appropriate and more naturalistic auditory context is crucial for understanding how the brain processes timbre. The studies mentioned thus far serve as stepping stones for designing further studies with higher ecological validity. It is important to note that real-world stimuli are complex, and disentangling neural correlates of timbre can be extremely challenging despite the considerably high ecological validity of the results obtained from them.

6.4 Neural Correlates of Global Sound

Context is crucial in uncovering the workings of the brain. In real life, we generally encounter combinations of sounds that result in complex emerging timbral soundscapes, for instance, that of a bustling market place or a busy highway. Furthermore, in the context of music, the co-occurrence of several instrument timbres results in an overall set of timbral relations, such as that of a jazz ensemble, a rock concert, or a symphony. This approach to timbre perception as a conglomerate of individual timbres has been largely neglected except in the context of mid-level representations, that is, sound textures that deal with naturally occurring environmental sounds. These sounds are considered to possess homogeneous properties that can be represented by summary statistics over time. However, in real-world settings, our brains are exposed to complex, temporally varying auditory scenes that require longer duration acoustic signals that are more ecologically valid. This section summarizes work performed in the naturalistic paradigm, especially in the context of music, and ends with a consideration of the challenges encountered and possible future research directions.

The global sound or the overall timbral mixture has been described as "polyphonic timbre" (Aucouturier 2006; Alluri and Toiviainen 2010). The term "polyphonic" should not be confused with the music theory term of polyphony (versus homophony or monophony); polyphonic describes the overall emerging sound that results in a mental construct (such as the set of timbres of a jazz ensemble). Global sound can be considered analogous to harmony in the context of pitch; in other words, pitch is to harmony as timbre is to global sound.

6.4.1 Naturalistic Auditory Paradigm: Encoding and Decoding Approaches

Alluri et al. (2012) were among the first researchers to look at timbre-feature processing in a naturalistic auditory paradigm in the context of music. Participants were asked to lie down in a scanner and listen to an 8-min Argentinian tango. Principal components were extracted from a set of low-level to mid-level audio descriptors related to timbre. Perceptual validation of the components obtained in this way resulted in higher levels of abstraction of timbre that represented holistic properties, such as *fullness* (spectral flux of the lower end of the spectrum) and *activity* (spectral flux of the middle to high end of the spectrum). Owing to the nature of fMRI signals (capturing slow-moving changes and not instantaneous ones), in the naturalistic paradigm these timbral components indeed would be representative of more global properties that lie on the higher end of abstraction. As a result, audio descriptor components that represented spectral flux in subbands of the spectrum (Alluri and Toiviainen 2010) possessed the highest positive correlations with activations in large areas of the temporal lobe (HG, MTG, and STG) in addition to the right rolandic operculum. Interhemispheric specialization in the auditory cortices was observed, specifically in the caudolateral and anterolateral parts of the STG; overall, the right temporal lobe displayed larger areas with significant correlations.

In addition, the results demonstrated negative correlations for the first time between timbre-related audio descriptors and activity in the default-mode network (DMN) areas of the cerebrum. The DMN is a neural circuit that constantly monitors the sensory environment and displays high activity during lack of focused attention on external events (McAvoy et al. 2008; Fox et al. 2009). As low values in descriptor components were mostly associated with sections in the stimulus with sparse texture played by the piano (thereby resulting in lower levels of auditory-cognitive load), the activation of the DMN during these moments is in line with previous results (Pallesen et al. 2009; Uddin et al. 2009). Interestingly, the DMN activity also was deactivated while processing the timbre of real-world materials such as plastic and metal (Lemaitre et al. 2017). On the other hand, high values in the timbral descriptors were associated with dense textures (high levels of auditory cognitive load), thereby causing a proportional reduction in activations in the DMN.

A subsequent follow-up study in the naturalistic paradigm with a different set of participants and stimuli revealed that the activations in the anterior STG could be best predicted by acoustic properties of the stimuli, which mostly comprised audio descriptors representing spectrotemporal modulations (Alluri et al. 2013). This finding lends further support to the conclusions of Giordano et al. (2013) that the right anterior STG is sensitive to spectral flux. In addition, activations in an important hub of the DMN (precuneus) could be predicted to a significant degree, thereby corroborating the notion of directed attention. Furthermore, a replication study of the original Alluri et al. (2012) study revealed high reliability in terms of the brain regions that were associated with processing timbral components (Burunat et al. 2016).

In a related decoding study, Toiviainen et al. (2014) reported the highest prediction accuracy for timbre-related audio descriptors from brain responses in comparison to rhythmical and tonal audio descriptors. The regions in the brain that contributed to the predictions encompassed STG, HG, rolandic operculum, and MTG with larger areas in the right hemisphere, thereby corroborating the results of the first study on the naturalistic paradigm (Alluri et al. 2012). Hoefle et al. (2018) applied an encoding-decoding approach to fMRI data collected in the naturalistic paradigm to understand complex auditory representations and to identify brain regions that contribute to the identification of musical pieces. Interestingly, the HG encoded low-level descriptors (associated with sound energy levels and smoothness of spectral energy distribution), whereas mid-level descriptors (associated with brightness) possessed representations in secondary auditory regions, including both anterior and posterior STG and the planum temporale and planum polare. This finding lends support to the possibility of interplay between the ventral and dorsal auditory pathways during timbre perception, as summarized by Bizley and Cohen (2013).

6.4.2 Decoding Global Sound Using Multivariate Pattern Analysis

Global sound is a significant perceptual component of music, especially in studies that involve tasks such as genre identification, categorization, or emotional affect attribution. Features representing global sound are vital in the design of computational systems that perform genre-based, style-based, and mood-based categorizations of music (Liu et al. 2003; Barbedo and Lopes 2007).

Several high-resolution (7 Tesla) fMRI studies have investigated musical genre representation in the brain. Casey et al. (2011) demonstrated via an MVPA approach that musical genre discrimination (among 6-s excerpts of ambient, rock and roll, heavy metal, classical, and country music) is attributed to a distributed population code in bilateral STS. Audio descriptors based on cepstral coefficients (see Caetano, Saitis, and Siedenburg, Chap. 11) were the best predictors of the voxel patterns of the STS. A subsequent decoding study by the same research group using the same dataset revealed that multivoxel patterns in the temporal cortex provided high genre classification accuracy in addition to stimulus identification at above-chance levels (Casey, 2017), thereby lending further support to a hierarchical model of timbre encoding.

6.5 Summary

To conclude, the several stages of timbre processing are best described by a hierarchical approach from low-level decompositions to mid-level summary statistics to high-level categorical representations in the brain. While the primary auditory cortex A1 is associated with low-level representations, the secondary regions, including STG and STS, serve as mid-level representations of auditory sensations. There exist dual pathways that represent the "where" and "what" of the sound on hand. Since timbre is associated with what is being heard, most studies have supported the key role of the ventral stream for processing various levels of timbre abstraction, which vary as a function of distance from A1 in the anterior direction. However, recent evidence suggests that information abstraction extends both in anterior and posterior directions starting from A1, and further research is needed. The auditory cortex has evolved to yield larger neural responses to relevant and natural stimuli; therefore, care is required in the interpretation of timbre processing of synthesized and unnatural sounds. Studies on high-level timbre representations in the context of music and speech report overwhelming evidence of secondary auditory regions with some reporting a right hemispheric bias (at least for music).

Intuitively, one could assume that a multidimensional attribute such as timbre would be represented in the brain not by independent isolated modules but over distributed neuronal populations that are tuned to process temporal, spectral, and combined spectrotemporal information (Santoro et al. 2014). Advances in machine-learning techniques, especially artificial neural networks, have given rise to the subfield of *deep learning*, which is based on algorithms and approaches inspired by the structure and functioning of the brain. One such deep neural network is an autoencoder, which learns to produce the same output as a given input through the creation of several levels of abstraction by forcing the input (e.g., the sound stimulus) to go through several layers with reduced neurons to arrive at a bottleneck. This approach is akin to mapping the spectrograms onto concise abstract representations. Comparing the patterns arising at each abstraction layer with specific regions in the brain, using MVPA or representation similarity analysis, would aid in zeroing in on the brain regions associated with various levels of abstraction of timbre.

The increasing popularity of multimodal fusion approaches (e.g., MEG+EEG, fMRI+EEG) appear to be attractive options to address timbre encoding from low to high levels of abstraction in the brain with fine-grained temporal and spectral resolution. Finally, with the increasing popularity of naturalistic listening situations and paradigms in neuroimaging, it is time to move on from controlled settings and univariate methods and address timbre processing in more realistic and holistic contexts.

Acknowledgements The authors would like to thank Prof. Stephen McAdams and Kai Siedenburg for their valuable inputs. V. A. extends her sincere gratitude to Prof. Elvira Brattico for opening the door to neuroscience and Prof. Petri Toiviainen for being the wonderful mentor that he is and providing never-ending support both professionally and personally.

Compliance with Ethics Requirements Vinoo Alluri declares that she has no conflict of interest.

Sudarsana Reddy Kadiri declares that he has no conflict of interest.

References

Allen EJ, Oxenham AJ (2014) Symmetric interactions and interference between pitch and timbre. J Acoust Soc Am 55455:1371–1379. https://doi.org/10.1121/1.4863269

Allen EJ, Burton PC, Olman CA, Oxenham AJ (2017) Representations of pitch and timbre variation in human auditory cortex. J Neurosci 37:1284–1293. https://doi.org/10.1523/JNEUROSCI.2336-16.2016

Alluri V, Toiviainen P (2010) Exploring perceptual and acoustical correlates of polyphonic timbre. Music Percept 27:223–242. https://doi.org/10.1525/mp.2010.27.3.223

Alluri V, Toiviainen P, Jääskeläinen IP et al (2012) Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. NeuroImage 59:3677–3689. https://doi.org/10.1016/j.neuroimage.2011.11.019

Alluri V, Toiviainen P, Lund TE et al (2013) From Vivaldi to Beatles and back: predicting lateralized brain responses to music. NeuroImage 83:627–636. https://doi.org/10.1016/j.neuroimage.2013.06.064

Aucouturier JJ (2006) Dix expériences sur la modélisation du timbre polyphonique (Ten experiments on the modeling of polyphonic timbre). Dissertation, Université Paris 6

Auzou P, Eustache F, Etevenon P et al (1995) Topographic EEG activations during timbre and pitch discrimination tasks using musical sounds. Neuropsychologia 33:25–37. https://doi.org/10.1016/0028-3932(94)00100-4

Baumann S, Griffiths TD, Sun L et al (2011) Orthogonal representation of sound dimensions in the primate midbrain. Nat Neurosci 14:423–425. https://doi.org/10.1038/nn.2771

Baumann S, Joly O, Rees A et al (2015) The topography of frequency and time representation in primate auditory cortices. elife 4:e03256. https://doi.org/10.7554/eLife.03256

Barbedo JGA, Lopes A (2007) Automatic genre classification of musical signals. EURASIP J Adv Signal Process, 2007(1): 157–168

Beauchamp MS, Lee K, Argall B, Martin A (2004) Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41:809–823. https://doi.org/10.1016/S0896-6273(04)00070-4

Belin P, Zatorre RJ, Lafaille P et al (2000) Voice-selective areas in human auditory cortex. Nature 403:309–312. https://doi.org/10.1038/35002078

Bizley JK, Cohen YE (2013) The what, where and how of auditory-object perception. Nat Publ Gr 14:693–707. https://doi.org/10.1038/nrn3565

Bizley JK, Walker KMM, Silverman BW et al (2009) Interdependent encoding of pitch, timbre, and spatial location in auditory cortex. J Neurosci 29:2064–2075. https://doi.org/10.1523/JNEUROSCI.4755-08.2009

Bregman AS (1990) Auditory scene analysis: the perceptual organization of sound. Cambridge (MA), MIT Press

Burunat I, Toiviainen P, Alluri V et al (2016) The reliability of continuous brain responses during naturalistic listening to music. NeuroImage 124:224–231. https://doi.org/10.1016/j.neuroimage.2015.09.005

Caclin A, Brattico E, Tervaniemi M et al (2006) Separate neural processing of timbre dimensions in auditory sensory memory. J Cogn Neurosci 18:1959–1972. https://doi.org/10.1162/jocn.2006.18.12.1959

Caclin A, McAdams S, Smith BK, Giard M-H (2008) Interactive processing of timbre dimensions: an exploration with event-related potentials. J Cogn Neurosci 20:49–64. https://doi.org/10.1162/jocn.2008.20.1.49

Casey MA (2017) Music of the 7Ts: predicting and decoding multivoxel fMRI responses with acoustic, schematic, and categorical music features. Front Psychol 8:1179. https://doi.org/10.3389/fpsyg.2017.01179

Casey M, Thompson J, Kang O, et al (2011) Population codes representing musical timbre for high-level fMRI categorization of music genres. In: Langs G, Rish I, Grosse-Wentrup M, Murphy B (eds) Machine learning and interpretation in neuroimaging. International workshop, MLINI 2011, held at NIPS 2011, Sierra Nevada, December 2011. Springer, Heidelberg, p 34–41

Crummer GC, Walton JP, Wayman JW et al (1994) Neural processing of musical timbre by musicians, nonmusicians, and musicians possessing absolute pitch. J Acoust Soc Am 95:2720–2727. https://doi.org/10.1121/1.409840

Davis T, LaRocque KF, Mumford JA et al (2014) What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. NeuroImage 97:271–283. https://doi.org/10.1016/j.neuroimage.2014.04.037

Deike S, Gaschler-Markefski B, Brechmann A, Scheich H (2004) Auditory stream segregation relying on timbre involves left auditory cortex. Neuroreport 15:1511–1514. https://doi.org/10.1097/01.wnr.0000132919.12990.34

Doehrmann O, Naumer MJ, Volz S et al (2008) Probing category selectivity for environmental sounds in the human auditory brain. Neuropsychologia 46:2776–2786. https://doi.org/10.1016/j.neuropsychologia.2008.05.011

Engelien A, Tüscher O, Hermans W, Isenberg N, Eidelberg D, Frith C et al (2006) Functional neuroanatomy of non-verbal semantic sound processing in humans. J Neural Transm 113:599–608

Ferrer R (2011) Timbral environments: an ecological approach to the cognition of timbre. Empir Musicol Rev 6:64–74

Formisano E, Kim DS, Di SF (2003) Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron 40:859–869

Formisano E, De Martino F, Bonte M, Goebel R (2008) “Who” is saying “what”? Brain-based decoding of human voice and speech. Science 322:970–973

Fox MD, Zhang D, Snyder AZ, Raichle ME (2009) The global signal and observed anticorrelated resting state brain networks. J Neurophys 101(6):3270–3283

Giordano BL, Rocchesso D, McAdams S (2010) Integration of acoustical information in the perception of impacted sound sources: the role of information accuracy and exploitability. J Exp Psychol Hum Percept Perform 36:462–476. https://doi.org/10.1037/a0018388

Giordano BL, McAdams S, Zatorre RJ et al (2013) Abstract encoding of auditory objects in cortical activity patterns. Cereb Cortex 23:2025–2037. https://doi.org/10.1093/cercor/bhs162

Giordano BL, Pernet C, Charest I et al (2014) Automatic domain-general processing of sound source identity in the left posterior middle frontal gyrus. Cortex 58:170–185. https://doi.org/10.1016/j.cortex.2014.06.005

Gjerdingen RO, Perrott D (2008) Scanning the dial: the rapid recognition of music genres. J New Music Res 37:93–100. https://doi.org/10.1080/09298210802479268

Goldstone RL (1998) Perceptual learning. Annu Rev Psychol 49:585–612. https://doi.org/10.1146/annurev.psych.49.1.585

Halpern AR, Zatorre RJ, Bouffard M, Johnson JA (2004) Behavioral and neural correlates of perceived and imagined musical timbre. Neuropsychologia 42:1281–1292. https://doi.org/10.1016/j.neuropsychologia.2003.12.017

Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R (2004) Intersubject synchronization of cortical activity during natural vision. Science 303:1634–1640

Hoefle S, Engel A, Basilio R et al (2018) Identifying musical pieces from fMRI data using encoding and decoding models. Sci Rep 8:2266. https://doi.org/10.1038/s41598-018-20732-3

Humphries C, Liebenthal E, Binder JR (2010) Tonotopic organization of human auditory cortex. NeuroImage 50:1202–1211. https://doi.org/10.1016/j.neuroimage.2010.01.046

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868. https://doi.org/10.1073/pnas.0600244103

Kumar S, Stephan KE, Warren JD et al (2007) Hierarchical processing of auditory objects in humans. PLoS Comput Biol 3:0977–0985. https://doi.org/10.1371/journal.pcbi.0030100

Langner G (2009) A map of periodicity orthogonal to frequency representation in the cat auditory cortex. Front Integr Neurosci 3:27. https://doi.org/10.3389/neuro.07.027.2009

Leaver AM, Rauschecker JP (2010) Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci 30:7604–7612. https://doi.org/10.1523/JNEUROSCI.0296-10.2010

Lemaitre G, Pyles JA, Halpern AR et al (2017) Who's that knocking at my door? Neural bases of sound source identification. Cereb Cortex:1–14. https://doi.org/10.1093/cercor/bhw397

Liu D, Lu L, Zhang HJ (2003) Automatic mood detection from acoustic music data. In: Hoos HH, Bainbridge D (eds) Proceedings of the 4th international conference on music information retrieval, October 2013. Johns Hopkins University, Baltimore, p 81–87

Lewis JW (2005) Distinct cortical pathways for processing tool versus animal sounds. J Neurosci 25:5148–5158. https://doi.org/10.1523/JNEUROSCI.0419-05.2005

Lewis JW, Phinney RE, Brefczynski-Lewis JA, DeYoe EA (2006) Lefties get it "right" when hearing tool sounds. J Cogn Neurosci 18:1314–1330. https://doi.org/10.1162/jocn.2006.18.8.1314

McAdams S, Winsberg S, Donnadieu S et al (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res 58:177–192. https://doi.org/10.1007/BF00419633

McAvoy M, Larson-Prior L, Nolan TS et al (2008) Resting states affect spontaneous bold oscillations in sensory and paralimbic cortex. J Neurpophysiol 100(2):922–931

McDermott JH, Simoncelli EP (2011) Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron 71:926–940. https://doi.org/10.1016/j.neuron.2011.06.032

McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. Nat Neurosci 16(4):493–498

Menon V, Levitin DJ, Smith BK et al (2002) Neural correlates of timbre change in harmonic sounds. NeuroImage 17:1742–1754. https://doi.org/10.1006/nimg.2002.1295

Miller LM, Escabí MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87:516–527. https://doi.org/10.1152/jn.00395.2001

Mlynarski W, McDermott JH (2018) Learning midlevel auditory codes from natural sound statistics. Neural Comput 30:631–669

Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88:1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Obleser J, Scott SK, Eulitz C (2006) Now you hear it, now you don't: Tran- sient traces of consonants and their nonspeech analogues in the human brain. Cereb Cortex 16:1069–1076

Pallesen KJ, Brattico E, Bailey CJ, Korvenoja A, Gjedde A (2009) Cognitive and emotional modulation of brain default operation. J. Cogn. Neurosci. 21(6), 901 1065–1080. https://doi.org/10.1162/jocn.2009.21086

Pantev C, Engelien A, Candia V, Elbert T (2001) Representational cortex in musicians. Plastic alterations in response to musical practice. Ann N Y Acad Sci 930:300–314. https://doi.org/10.1111/j.1749-6632.2001.tb05740.x

Petersen S, Posner M (2012) The attention system of the human brain: 20 years after. Annu Rev Neurosci 21:73–89. https://doi.org/10.1146/annurev-neuro-062111-150525.The

Platel H, Price C, Baron JC, et al (1997) The structural components of music perception. A functional anatomical study. Brain 120 (Pt 2:229–243. doi: https://doi.org/10.1093/brain/120.2.229

Rauschecker JP (1998) Cortical processing of complex sounds. Curr Opin Neurobiol 8(4):516–521. https://doi.org/10.1016/S0959-4388(98)80040-8

Rauschecker JP (2012) Ventral and dorsal streams in the evolution of speech and language. Front Evol Neurosci 4:7. https://doi.org/10.3389/fnevo.2012.00007

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718–724. https://doi.org/10.1038/nn.2331

Ress D, Chandrasekaran B (2013) Tonotopic organization in the depth of human inferior colliculus. Front Hum Neurosci 7:586. https://doi.org/10.3389/fnhum.2013.00586

Rodriguez FA, Read HL, Escabi MA (2010) Spectral and temporal modulation tradeoff in the inferior colliculus. J Neurophysiol 103:887–903. https://doi.org/10.1152/jn.00813.2009

Samson S (2003) Neuropsychological studies of musical timbre. Ann N Y Acad Sci 999:144–151

Samson S, Zatorre RJ, Ramsay JO (2002) Deficits of musical timbre perception after unilateral temporal-lobe lesion revealed with multidimensional scaling. Brain 125:511–523. https://doi.org/10.1093/brain/awf051

Samson F, Zeffiro TA, Toussaint A, Belin P (2011) Stimulus complexity and categorical effects in human auditory cortex: an activation likelihood estimation meta-analysis. Front Psychol 1:241. https://doi.org/10.3389/fpsyg.2010.00241

Sandell GJ (1998) Macrotimbre: Contribution of attack and steady state, Proc. 16th Int. Congress on Acoustics and 135th Meeting of the Acoustical Society of America, Vol. 3, Seattle (Acoustical Society of America, Woodbury, NY), pp 1881–1882

Santoro R, Moerel M, De Martino F et al (2014) Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput Biol 10(1):e1003412. https://doi.org/10.1371/journal.pcbi.1003412

Staeren N, Renvall H, De Martino F et al (2009) Sound categories are represented as distributed patterns in the human auditory cortex. Curr Biol 19:498–502. https://doi.org/10.1016/j.cub.2009.01.066

Stelzer J, Chen Y, Turner R (2013) Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. NeuroImage 65:69–82. https://doi.org/10.1016/j.neuroimage.2012.09.063

Theunissen FE, Elie JE (2014) Neural processing of natural sounds. Nat Rev Neurosci 15:355–366

Toiviainen P, Tervaniemi M, Louhivuori J et al (1998) Timbre similarity: convergence of neural, behavioral, and computational approaches. Music Percept 16:223–241. https://doi.org/10.2307/40285788

Toiviainen P, Alluri V, Brattico E et al (2014) Capturing the musical brain with lasso: dynamic decoding of musical features from fMRI data. NeuroImage 88:170–180

Town SM, Bizley JK (2013) Neural and behavioral investigations into timbre perception. Front Syst Neurosci 7:1–14. https://doi.org/10.3389/fnsys.2013.00088

Uddin LQ, Clare Kelly A, Biswal BB, Xavier Castellanos F, Milham MP (2009) Functional connectivity of default mode network components: Correlation, anticorrelation, and causality. Hum. Brain Mapp., 30: 625–637. https://doi.org/10.1002/hbm.20531

Varoquaux G, Thirion B (2014) How machine learning is shaping cognitive neuroimaging. Gigascience 3:28

Warren JD, Jennings AR, Griffiths TD (2005) Analysis of the spectral envelope of sounds by the human brain. NeuroImage 24:1052–1057. https://doi.org/10.1016/j.neuroimage.2004.10.031

Wessinger CM, Vanmeter J, Tian B et al (2001) Hierarchical organization of the Human Auditory Cortex Revealed by functional magnetic resonance imaging. J Cogn Neurosci 13(1):1–7

Zatorre RJ, Belin P (2001) Spectral and temporal processing in human auditory cortex. Cereb Cortex 11:946–953. https://doi.org/10.1093/cercor/11.10.94

Part II
Specific Scenarios

Chapter 7
Voice Processing and Voice-Identity Recognition

Samuel Robert Mathias and Katharina von Kriegstein

Abstract The human voice is the most important sound source in our environment, not only because it produces speech, but also because it conveys information about the speaker. In many situations, listeners understand the speech message and recognize the speaker with minimal effort. Psychophysical studies have investigated which voice qualities (such as vocal timbre) distinguish speakers and allow listeners to recognize speakers. Glottal and vocal tract characteristics strongly influence perceived similarity between speakers and serve as cues for voice-identity recognition. However, the importance of a particular voice quality for voice-identity recognition depends on the speaker and the stimulus. Voice-identity recognition relies on a network of brain regions comprising a core system of auditory regions within the temporal lobe (including regions dedicated to processing glottal and vocal tract characteristics and regions that play more abstract roles) and an extended system of nonauditory regions representing information associated with specific voice identities (e.g., faces and names). This brain network is supported by early, direct connections between the core voice system and an analogous core face system. Precisely how all these brain regions work together to accomplish voice-identity recognition remains an open question; answering it will require rigorous testing of hypotheses derived from theoretical accounts of voice processing.

Keywords Congenital phonagnosia · Core face system · Core voice system
Glottal-pulse rate · Vocal recognition · Vocal timbre · Vocal tract length

S. R. Mathias (✉)
Neurocognition, Neurocomputation and Neurogenetics Division, Yale University
School of Medicine, New Haven, CT, USA
e-mail: samuel.mathias@yale.edu

K. von Kriegstein
Technische Universität Dresden, Dresden, Germany

Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
e-mail: katharina.von_kriegstein@tu-dresden.de

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_7

Abbreviations

a	anterior
BOLD	blood-oxygen-level-dependent
d	distance measure
FFA	fusiform face area
fMRI	functional magnetic resonance imaging
FRU	facial recognition units
GPR	glottal-pulse rate
HG	Hechl's gyrus
HNR	harmonics-to-noise ratio
IFG	inferior frontal gyrus
IPL	inferior parietal lobe
JND	just noticeable difference
M	middle
MEG	magnetoencephalography
P	posterior
PIN	person-identity nodes
PT	planum temporale
STG	superior temporal gyrus
STS	superior temporal sulcus
Th	perceptual threshold
TVA	temporal voice areas
VLPFC	ventrolateral prefrontal cortex
VRU	voice recognition units
VTL	vocal-tract length

7.1 Introduction

The human voice is arguably the most important sound source in our environment. In addition to producing speech, it conveys a wealth of information about the speaker, including their sex, approximate age and body size, place of origin, and current emotional state. Unsurprisingly, the human brain is highly specialized in voice processing. This specialization allows normal-hearing listeners to understand the speech message, determine many characteristics of the speaker, and recognize a personally familiar or famous speaker, all at the same time. In fact, humans have routinely outperformed computer algorithms at both speech and voice-identity recognition, particularly under suboptimal listening conditions (Kitaoka et al. 2014; Hautamäki et al. 2015; but also see Kell et al. 2018).

This chapter discusses the known and unknown aspects of human voice processing and voice-identity recognition. Section 7.2 considers psychophysical studies on the acoustic and perceptual differences between voices and how listeners recognize who is speaking. Section 7.3 involves a discussion of brain regions exhibiting voice sensitivity (or selectivity), regions involved in processing familiar and unfamiliar voices, regions involved in voice-identity recognition specifically, and disorders of voice processing. Section 7.4 discusses theoretical models of voice-identity recognition and evaluates whether such models are consistent with current evidence.

The chapter has a strong emphasis on voice processing and voice-identity recognition involving speech rather than other kinds of vocal sounds (e.g., coughs, cries, laughter). While nonspeech vocal sounds obviously carry important information, including information about voice identity, speech is the most common class of vocal sound, and voices are most easily recognized from speech. This focus on speech also emphasizes the important point that speech-message and voice-identity recognition are not independent. Not only do they rarely occur in isolation, but listeners routinely use voice-identity information to improve speech recognition and vice versa (Wester 2012; Kreitewolf et al. 2017); this point is sometimes overlooked by theoretical accounts of voice processing.

7.2 Psychophysics of Voice Processing

7.2.1 Differences Between Voices

There are enormous acoustic differences between speech sounds produced by different speakers, even when the linguistic content of those sounds is identical. The classic study by Peterson and Barney (1952) illustrates some of these differences. The authors measured the fundamental frequencies (f_0) and the frequencies and amplitudes of the first three vowel formants (f_{1-3}) from recordings of many speakers, including men, women, and children, mostly from the Middle Atlantic region of the USA, reading monosyllabic words of the kind /hVd/ (e.g., heed, hid, head). They found that the same vowel produced by two different speakers could be nothing alike acoustically. For instance, one speaker's "hid" could be more similar to another speaker's "head" (in terms of the absolute frequencies of its formants) than to examples from its own linguistic category (Fig. 7.1). Despite these large acoustic differences, listeners were able to correctly recognize approximately 95% of the words, highlighting an intriguing conceptual problem: How do listeners effortlessly overcome such large acoustic differences to discern so much information from speech?

Most previous research has approached this conceptual problem from the perspective of speech perception, attempting to understand how listeners anticipate and correct for speaker-related acoustic differences when recognizing the linguistic content of speech. This process is called *speaker normalization* (Johnson 2005; von Kriegstein et al. 2010). However, the fact that listeners often discriminate and recognize different speakers just as easily as they understand speech implies that the opposite kind of normalization (*speech normalization*) occurs as well. Put another way, listeners must be able to perceive voice features that are constant across speech

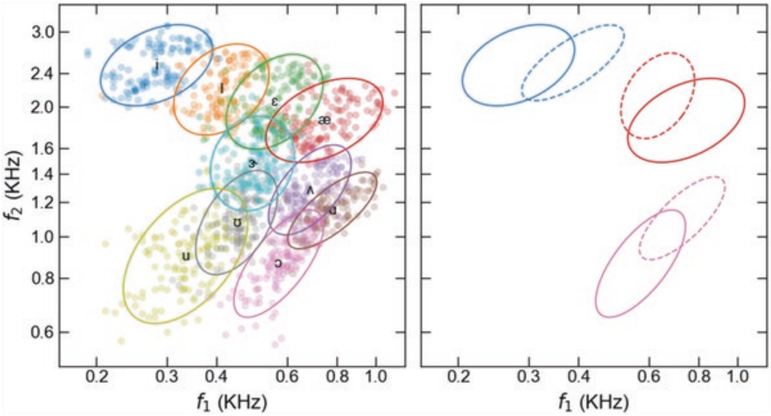

Fig. 7.1 Variations in spoken vowels. *Left panel*: f_1 and f_2 measurements from ten vowels spoken by a set of Middle-Atlantic American speakers from a study by Peterson and Barney (1952). These famous data highlight the extreme acoustic variability between recordings of the same vowel spoken by different speakers. *Right panel*: *solid lines* are three-sigma ellipses from three of the vowels recorded by Peterson and Barney; *dashed lines* are three-sigma ellipses from the same three vowels recorded from speakers (mostly from southern Michigan) in a later study by Hillenbrand et al. (1995). The striking differences between the two data sets were likely caused by differences in the speaking styles of the two groups, highlighting the importance of largely nonanatomical factors to the perception of voice quality. (Data from Peterson and Barney 1952; right panel from Hillenbrand et al. 1995; used with permission)

sounds produced by the same speaker yet different across sounds produced by different speakers.

The set of features that makes a particular speaker's voice unique is called their *voice quality* (Kreiman et al. 2005). It could also be called *vocal timbre*. The former term is used more commonly than the latter term, probably because the former implies that voices may be ordered (from pathological to modal quality). This ordering can be useful in the assessment and treatment of articulation difficulties (see Sect. 7.2.2.2 for a brief discussion of voice-quality schemes that are used to assess such pathologies) (also see Saitis and Weinzierl, Chap. 5). Furthermore, voice quality is the more general term since it can include vocal pitch (Sect. 7.2.2.1).

Over the years, there have been many attempts to define and measure voice quality with mixed results. *Glottal-pulse rate* (GPR) and *vocal tract length* (VTL) have emerged as two obvious dimensions of voice quality. The GPR of the voice determines its perceived pitch, while the acoustic effect of VTL is an important aspect of vocal timbre. Both GPR and VTL have clear anatomical antecedents (Fig. 7.2), are acoustically and perceptually distinct from one another, and listeners routinely use them to determine certain characteristics of the speaker (e.g., sex and relative size) (Lavner et al. 2000). However, there has been little success in identifying other reliable dimensions of voice quality beyond GPR and VTL.

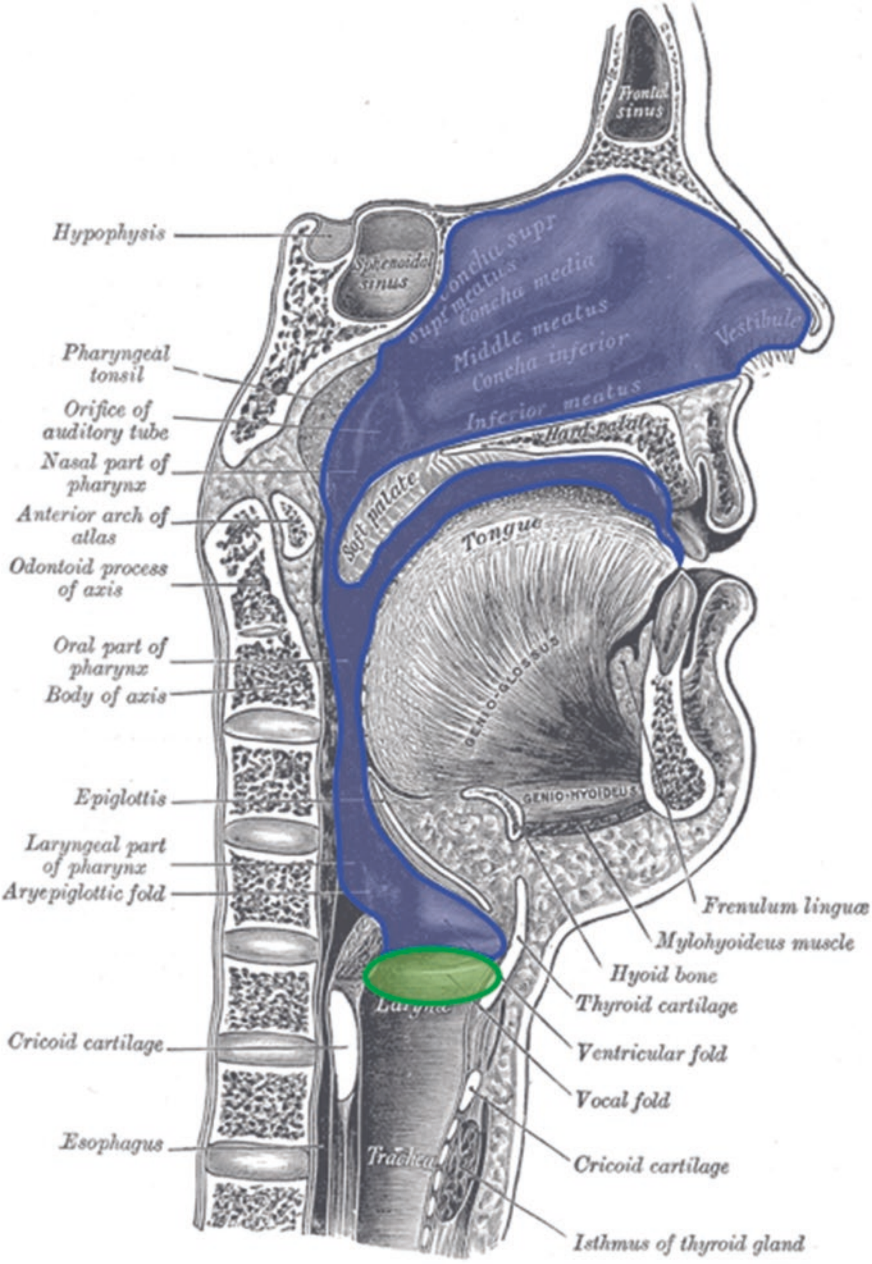

Fig. 7.2 Sagittal section through a human head and neck with the vocal folds highlighted in *green oval* and the vocal tract highlighted in *blue*, the anatomical antecedents of glottal-pulse rate and vocal tract length, respectively. (Adapted from Gray 1918, Mathias and von Kriegstein 2014)

7.2.2 Glottal-Pulse Rate

Many vocal sounds are *voiced*, meaning that they are produced by vibration of the vocal folds. The periodic waveform created by this vibration is called the *glottal pulse*. The rate of glottal vibration, or GPR, determines the speaker's f_0. Speakers routinely modulate their GPR to distinguish between statements and questions in nontonal languages or to convey meaning in tonal languages. However, a speaker's long-term average GPR is relatively stable. The GPR is perceived as pitch (for a review of the psychophysics of pitch perception, see Plack and Oxenham 2005) and is by far the most consistent predictor of perceived similarity between speakers. In a representative study, Baumann and Belin (2010) presented listeners with pairs of French vowels and instructed them to rate how likely they thought it was that the same person spoke both vowels. The vowels could be the same or different and could be produced by the same or different speakers (speakers within a pair were always of the same sex). The speakers were all unfamiliar to the listeners (see Sects. 7.2.2.1 and 7.2.2.2). Using multidimensional scaling, the authors found that the primary dimension along which listeners based their speaker similarity judgements closely corresponded to the mean f_0 of the vowels (i.e., GPR). This basic finding has been a consistent feature of studies of this kind, regardless of the type of speech sounds used (isolated vowels, whole words, or sentences) and whether the speakers were men, women, or a mixture of both.

The GPR is an obvious cue for determining a speaker's sex. Typical f_0 values are 120 Hz for men and 210 Hz for women, amounting to an average difference of around 10 semitones or a minor seventh. By way of comparison, most listeners can easily discriminate f_0 differences between synthesized vowels that are smaller than half a semitone (Smith et al. 2005). This difference is because, on average, men have longer and thicker vocal folds that vibrate more slowly than those in women (Titze 1989). While GPR is the strongest cue for speaker sex, experiments using synthesized speech have shown that GPR alone is not enough to guarantee robust perception of a speaker as either male or female. For example, Hillenbrand and Clark (2009) instructed listeners to judge the sex of speakers from sentences that were processed using a vocoder to shift either the f_0, the first two formants, or both f_0 and the formants into the range of speakers of the opposite sex. Shifting f_0 alone caused males to be perceived as females about 34% of the time and the opposite about 19% of the time. Shifting both f_0 and the formats caused perceived sex to shift more often but still not every time (about 82%), suggesting that there were residual perceptual cues to the sex of the speaker.

Glottal-pulse rate is somewhat predictive of other speaker characteristics. For example, there is a relationship between age, sex, and GPR, but it is not straightforward. While adults have slower GPRs than children due to the enlargement and thickening of the vocal folds during development, the rate of change is more dramatic during puberty in males than in females (Fouquet et al. 2016). Moreover, in older adulthood, GPR increases in men yet decreases in women (Hollien and

Shipp 1972; Stoicheff 1981). Environmental factors, such as smoking, also influence GPR (Gilbert and Weismer 1974). Perhaps surprisingly, GPR is not a good indicator of speaker size: after controlling for age and sex, GPR and speaker height/weight are essentially uncorrelated (Künzel 1989).

7.2.2.1 Vocal Tract Length

The supralaryngeal vocal tract and lips act as a filter on the glottal pulse, amplifying and attenuating certain frequencies to create formants. Speakers modify the shape of their vocal tract to produce different speech sounds, but their overall VTL is stable. Acoustically, the spectral envelope and formants of speech produced by a speaker with a long VTL are shifted downward in frequency relative to a speaker with a short VTL (Fig. 7.3), and VTL correlates very strongly with speaker height and/or weight ($r > 0.9$; Fitch and Giedd 1999). Several studies have shown that VTL is the primary perceptual cue for speaker size. For example, Smith and Patterson (2005) presented listeners with sustained vowels, all spoken by the same speaker. The vowels were vocoded to create versions that sounded as if they were spoken by speakers with different GPR and VTL combinations. Listeners judged the height of the speakers using a seven-point scale. When GPR/VTL values remained within the biologically plausible range, VTL was the dominant influence on a listener's height judgements. Other similar studies have shown that listeners are highly sensitive to VTL modifications (Smith et al. 2005).

Vocal tract length also influences the perceived similarity of voices. Previous multidimensional-scaling studies (Murray and Singh 1980; Baumann and Belin 2010) found that listeners' similarity judgements of unfamiliar speakers' voices were influenced by the speaker's mean formant frequencies, which are determined partly by VTL, although this influence was generally weaker than that of GPR. Unfortunately, it is not possible to determine the size of the effect of VTL per se on the listeners' similarity judgements from these studies because none of them measured their speakers' VTLs or used vocoded stimuli. However, in another relevant study, Gaudrain et al. (2009) presented listeners with pairs of three-vowel sequences, and asked them to judge whether "[it was] possible that both sequences were uttered by the same speaker?" The same male speaker spoke all the vowels, which were vocoded so that his GPR and/or VTL differed between sequences. Listeners reported hearing different speakers when there was a between-sequence VTL difference of around 25%, or 2.6 times their just-noticeable difference (JND) for VTL. Listeners reported hearing different speakers when there was a between-sequence GPR difference of around 45%, or 11.1 times their JND for GPR. Therefore, according to this study, perceptually smaller differences in VTL are required to elicit the perception of a different talker than those of GPR, suggesting that VTL is more important than GPR to the perceived similarity of voices, at least when listening to isolated vowels.

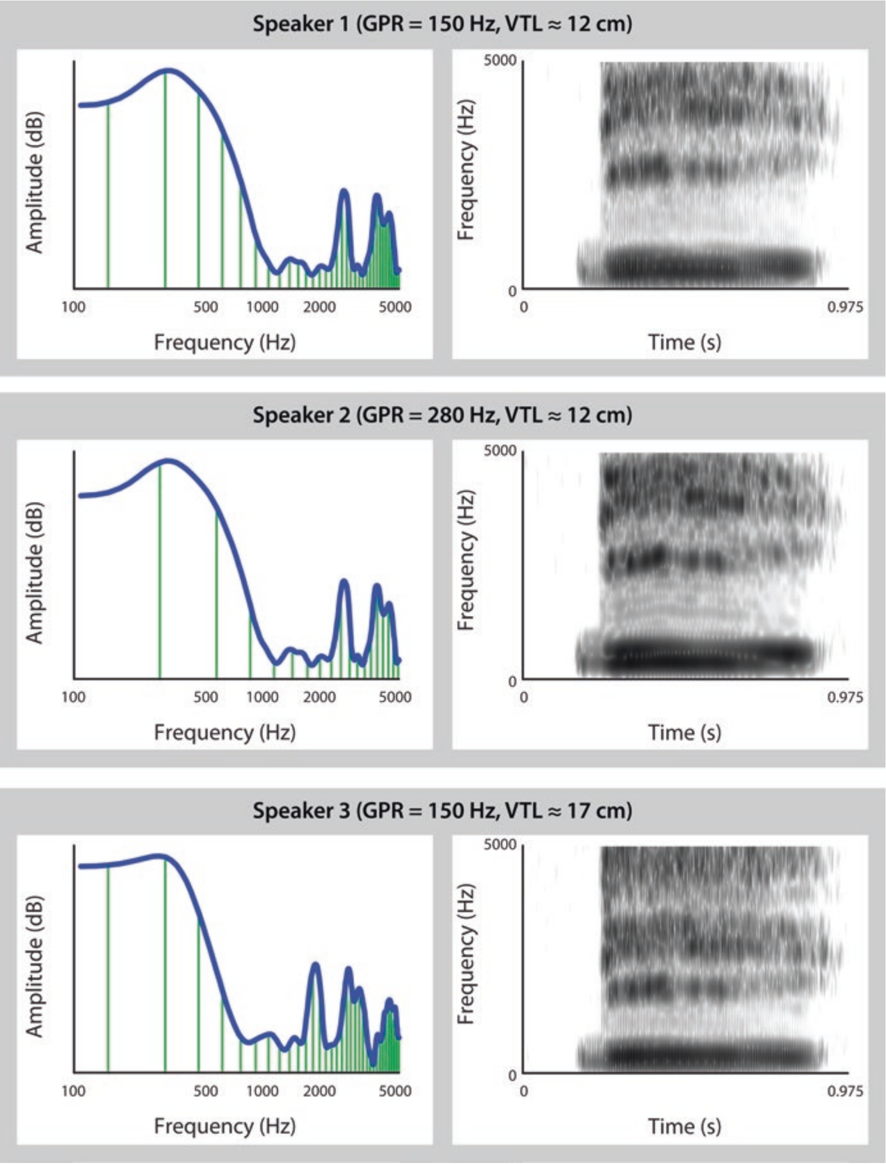

Fig. 7.3 Each row shows a cartoon spectral profile (*left*) and actual spectrogram (*right*) of an elongated /i/ vowel spoken by the same speaker but vocoded to have very different glottal-pulse rates (*GPR*s) and vocal tract length (*VTL*) values. The *middle vowel* has a faster GPR than the *top vowel*, and the *bottom vowel* has a longer VTL than the *top vowel*. In the *left panels*, one can see how GPR controls the f_0 and harmonics (*green vertical lines*), while VTL controls the formants (*blue curves*). In the *right panels*, the formants are clearly visible as dark spectral regions. (Adapted from Mathias and von Kriegstein 2014)

7.2.2.2 Schemes of Voice Quality

It is commonplace to describe a speaker's voice using various descriptive terms such as breathy, creaky, hoarse, and a variety of other adjectives (see Saitis and Weinzierl, Chap. 5). There have been many attempts to codify such terms, probably the most successful of which was Laver's (1980) scheme of *vocal-profile analysis*. Based on an extensive review of the pre-1980's literature, Laver described a number of *articulatory settings* or deviations from a typical voice. Each setting had an anatomical antecedent. For instance, breathiness was defined by low tension in the vocal chords. Other glottal settings included whisperiness, creakiness, harshness, and falsetto. Settings related to the vocal tract included lip configuration, laryngeal position, jaw position, and tongue position. Wherever possible, Laver provided examples of the acoustic and perceptual consequences of various settings, although these examples were not exhaustive.

Voice-quality schemes, such as the one proposed by Laver (1980), have several limitations. For instance, it is difficult to make the descriptions of settings precise enough to allow them to be measured acoustically. Instead, different settings are usually measured using the subjective ratings of phoneticians, which can have poor inter-rater reliability (Kreiman and Gerratt 1998). If listeners cannot agree on a speaker's settings, their utility for characterizing voice quality is questionable. Another limitation of voice-quality schemes is that they largely ignore factors unrelated to anatomy, including language, accent, dialect, and idiolect (personal speaking style). Such differences, though perhaps difficult to define and measure, are likely to be important for recognizing familiar voices when listening to complex stimuli such as whole sentences (see Sect. 7.2.2.2). Even isolated vowels are influenced greatly by speaking style as demonstrated by replications of Peterson and Barney's (1952) study using speakers from different regions within the USA (Fig. 7.1).

7.2.3 Voice-Identity Recognition

All of the behavioral studies discussed in Sect. 7.2.1 involved *discrimination*, requiring listeners to compare, judge, or rate voices in terms of their perceptual features, usually without prior exposure to the speakers or stimuli. By contrast, *recognition* requires listeners to identify speakers they heard previously. It does not necessarily follow that the acoustic or perceptual cues relevant for discrimination are the same as those relevant for recognition. Indeed, McAdams (Chap. 2) and Agus, Suied, and Pressnitzer (Chap. 3) discuss evidence of different acoustic and perceptual cues supporting nonvocal timbre discrimination versus recognition. Moreover, discrimination and recognition may rely on different kinds of perceptual and cognitive processes, and the neural correlates of voice discrimination and recognition might at least partly dissociate (see Sect. 7.3).

7.2.3.1 Unfamiliar Speakers

Many studies have investigated the circumstances that influence the reliability of *ear-witness testimony* or how well listeners can pick out a target speaker from an auditory line-up (for a review, see Yarmey 2007). These studies can be thought of as investigating the recognition of unfamiliar speakers, since listeners had limited exposure to them prior to testing. As might be expected, recognition rates improve monotonically with the duration of the stimuli and decline as a function of the retention interval. Listeners are best at recognition when the target speaker and the others in the line-up all have accents similar to their own (Stevenage et al. 2012) and when all of the speech is in their native language (Wester 2012). There has been some debate about whether recognition rates are influenced by the sex of the listener and/or speaker. A meta-analysis of several ear-witness experiments suggests that female listeners are significantly better at recognizing female than male speakers, but no similar advantage exists when male listeners hear male speakers (Wilding and Cook 2000).

One consistent finding from such studies is that recognition of an unfamiliar speaker is often very fragile. In many of these studies, listeners barely performed above chance levels. Unfamiliar-speaker recognition is also easily disrupted if the speaker deliberately disguises his or her voice between the initial and test stimuli. For example, even a relatively minor change in speaking style, such as switching from a normal to an angry tone, is enough to disrupt recognition (Saslove and Yarmey 1980).

7.2.3.2 Familiar Speakers

Familiar speakers are often personal acquaintances or famous speakers. Following Maguinness et al. (2018), we define *familiar voices* as those of speakers to whom the listener has had considerable prolonged exposure, either via social interactions (personally familiar speakers, Lavner et al. 2000; von Kriegstein and Giraud 2004) or the media (famous speakers, Van Lancker et al. 1985). Listeners typically know other information about a familiar speaker in addition to his/her voice, including his/her face, name, and biographical details (e.g., where and when they last saw, heard, or met the speaker). It is also possible to induce familiarity with a previously unfamiliar speaker to the point that listeners can reliably identify them from previously unheard speech recordings via training (Sheffert et al. 2002). However, laboratory-based training typically involves less exposure to speakers' voices and does not provide the rich supra-modal associations of speaker familiarization under natural conditions; thus, in most studies, laboratory-based familiarization is likely only a simulacrum of real world familiarization. We therefore call voices that have been familiarized by laboratory-based training *recently familiarized voices* (e.g., Sect. 7.3.2.1) (also see Maguinness et al. 2018).

Glottal and vocal tract characteristics appear to play a significant but modest role in familiar-speaker recognition. In perhaps the most extensive study of its kind, Lavner et al. (2000) instructed thirty listeners to identify twenty male speakers from modified recordings of /a/ vowels. At the time of assessment, all listeners and speakers lived in the same kibbutz, or commune, and had been living there for at least 5 years; therefore, they were highly familiar with one another's voices. Acoustic modifications included shifting the frequencies of individual formants or combinations of formants to those of another speaker, shifting the whole spectral envelope (approximately corresponding to VTL), changing the f_0 (i.e., GPR), changing the shape of the glottal waveform, and creating voices with the glottal characteristics of one speaker and the vocal tract characteristics of another. Shifting the formant frequencies (particularly the higher formants that remain relatively more stable than lower formants across speech produced by a single speaker) and shifting the whole spectral envelope had the largest effects on voice-identity recognition. Shifting the f_0 also had a strong effect, yet changing the shape of the glottal waveform had little impact on recognition. Whether these findings can be generalized to whole words or sentences is currently unclear.

In contrast to unfamiliar-speaker recognition, familiar-speaker recognition is often robust to extreme acoustic manipulation, especially when listening to stimuli with longer durations, such as whole sentences. For example, Remez et al. (1997) found that listeners could recognize some of their colleagues above chance levels from *sine-wave speech* (intelligible synthetic stimuli composed of three time-varying sinusoids that trace the frequency contours of the formants from a real sentence). Sine-wave speech contains very few, if any, traditional voice-quality cues—GPR and other glottal characteristics are lost completely, although VTL may be partially inferred—but retain some information regarding speaking style. In another study, Van Lancker et al. (1985) found that listeners could sometimes recognize famous speakers from time-reversed speech, which had the same long-term spectrotemporal properties as natural speech but little information about speaking style. In these two studies, the ability to recognize familiar speakers was reduced, but not eliminated, by manipulations that disrupted either spectrotemporal cues or cues related to speaking style. Taken together, these results suggest that listeners can use different types of cues for familiar-speaker recognition depending on the stimuli.

The relative importance of a particular cue for familiar-speaker recognition is speaker-dependent. For example, Lavner et al. (2000) found that some speakers were difficult to recognize from speech in which their vocal-tract characteristics were modified, but the same manipulations hardly affected the recognition of other speakers. In a follow-up study, the authors used multiple linear regressions to predict listeners' recognition scores from acoustic measurements. They found that regression weights for different predictors varied considerably across speakers (Lavner et al. 2001). Whether such cue variability exists when listeners recognize or discriminate unfamiliar speakers (Sect. 7.2.2.1) remains to be shown.

Taken together, these results suggest that listeners are able to draw from any number of potential cues for familiar-speaker recognition, and the relative importance of any given cue depends on the type of stimuli and the speaker. In other words, the set of cues for familiar-speaker recognition is probably neither ordered nor closed.

7.3 Neurobiology of Voice Processing

7.3.1 Voice Sensitivity and Selectivity

Understanding voice processing in the neurotypical human brain took a considerable leap forward at the turn of this century. Using functional magnetic resonance imagining (fMRI), Belin et al. (2000) contrasted the blood-oxygen-level-dependent (BOLD) response during passive listening to vocal and nonvocal sounds. Vocal sounds included speech and voiced nonspeech (singing, babble, cries, etc.) and nonvocal sounds included environmental sounds, animal vocalizations, and sounds from manmade objects. The authors found that regions located primarily along the upper bank of the superior temporal sulcus (STS) responded more strongly when listeners heard the vocal sounds. In other words, these regions exhibited *voice sensitivity*. Voice sensitivity was bilateral but appeared to be (slightly) stronger in the right hemisphere. There were also separate voice-sensitive maxima in the posterior, middle, and anterior portions of the sulcus.

The basic observation of voice sensitivity in the superior portions of the temporal lobe has been replicated many times. Recently, Pernet et al. (2015) analyzed fMRI data from 218 listeners who all completed the same voice-area localizer scan (a brief fMRI experiment used to define specific regions of interest within an individual's brain prior to the main experiment). The design of this localizer was very similar to the main experiment by Belin et al. (2000), which involved passive listening to voice and nonvoice sounds. Virtually all listeners exhibited voice sensitivity bilaterally in the superior temporal sulcus or gyrus (STS/STG). A group-level cluster analysis confirmed the findings of distinct posterior, middle, and anterior voice-sensitive peaks by Belin et al. (2000) (Fig. 7.4). Since the precise locations of these maxima tend to vary considerably across listeners—sometimes localized to the STS, the STG, or even to upper portions of the middle temporal gyrus—it is convenient to call them the *temporal voice areas* (TVAs) (e.g., von Kriegstein and Giraud 2006). Pernet and colleagues also detected voice sensitivity in regions outside the temporal lobes (extra-temporal voice areas), including parts of the inferior frontal gyrus (IFG), several brainstem structures, and portions of the thalamus and amygdala. The authors speculated that the extra-temporal areas were responsible for processing the linguistic or emotional content of the voice sounds. They might also be involved in voice-identity recognition (Zäske et al. 2017).

Recent advances in nonhuman fMRI have allowed researchers to identify brain regions that exhibit sensitivity to conspecific vocalizations in other species. In these studies, the comparison of responses to vocal versus nonvocal sounds showed evidence for evolutionary counterparts to human voice-sensitive regions in the monkey supratemporal plane (Petkov et al. 2008; Perrodin et al. 2011) and in the temporal lobe of dogs (Andics et al. 2014).

There is an important distinction between voice sensitivity and *voice selectivity*. Voice sensitivity is often used to describe brain regions that respond more strongly to vocal sounds than other stimuli. Brain regions can be voice sensitive for any

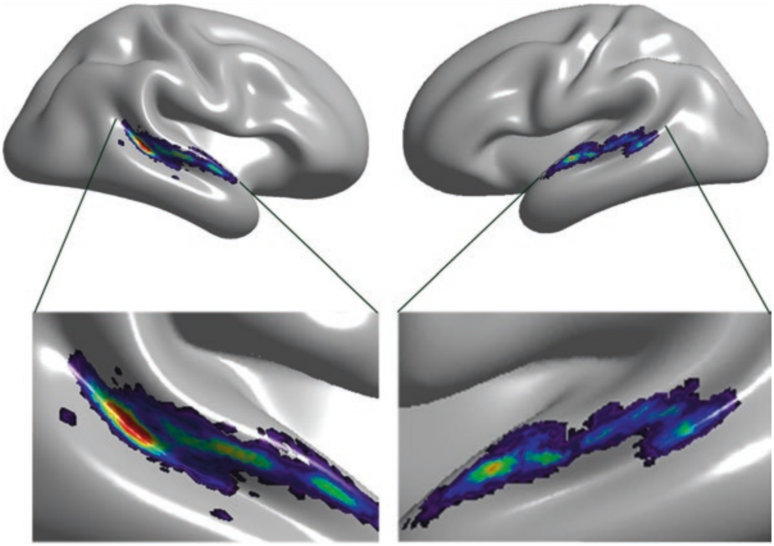

Fig. 7.4 Analysis of fMRI data from 218 listeners projected onto a default average brain. This analysis revealed three distinct voice-sensitive peaks along the superior temporal sulcus/gyrus in each hemisphere. These are commonly referred to as the temporal voice areas. *a*, anterior; *m*, middle; *p*, posterior; *TVA*, temporal voice area. (Taken from Pernet et al. 2015; used with permission)

number of reasons; for example, the region might be sensitive to acoustic or perceptual features more common in vocal than nonvocal sounds. By contrast, voice selectivity is often used to describe brain regions that respond more strongly to voices per se rather than the acoustic features of voices. Whether TVAs are voice selective in this strict sense or merely voice sensitive is a difficult question to answer. Consider the finding that TVAs in the human brain respond more strongly to human than animal vocalizations (Fecteau et al. 2004). On the one hand, this finding could be used to argue in favor of voice selectivity because human and animal vocalizations are acoustically similar in some respects, such as overall complexity and usually having clear and dynamic pitches. On the other hand, human and animal vocalizations obviously have other salient perceptual differences.

In another relevant study, Agus et al. (2017) presented listeners with natural voices and *auditory chimeras* (also see Agus, Suied, and Pressnitzer, Chap. 3). Chimeras are composites of two natural sounds, in this case one vocal and one nonvocal. The authors found that one region in the right temporal lobe responded more strongly to sequences of veridical vocal sounds than chimeras, even though both sequences contained many of the same acoustic features. Does this result mean that only right-lateralized (not left-lateralized) TVAs are voice selective? What about distinct TVAs within the same hemisphere (cf., Pernet et al. 2015)?

A particular brain region might be voice sensitive/selective because it performs a specific voice-related task, such as understanding the speech message, perceiving emotion, or recognizing voice identity. It is usually very difficult or impossible to disentangle these possibilities from fMRI studies involving passive listening to vocal and nonvocal stimuli. For example, Pernet et al. (2015) attempted to discern the functions of voice-sensitive regions by subcategorizing vocal sounds (neutral and nonintelligible, emotional and nonintelligible, and intelligible) and comparing responses across subcategories within each region. Almost every region responded most strongly to the intelligible sounds. However, in another study, Belin et al. (2002) contrasted nonspeech vocal sounds to spectrally scrambled versions of the same stimuli and found that only the right anterior TVA responded more strongly to the vocal sounds. As discussed in detail later (Sect. 7.3.2), subsequent studies combining stimulus and task manipulations have shed considerable light on the specific functionality of the TVAs.

Another important point is that a particular brain region may play an important role in voice processing without being voice sensitive/selective. For example, perception of GPR is important for linguistic processing and perceiving the sex of the speaker (Sect 7.2.2.1), but whether the brain has a region that is dedicated to processing vocal pitch per se is unclear. von Kriegstein et al. (2010) found that anterolateral Hechl's gyrus (HG) bilaterally responded more strongly to sequences of speech sounds from speakers with different GPRs than to sequences from speakers who all had the same GPR but different VTLs (see also Kreitewolf et al. 2014). These regions are commonly implicated in studies of nonvocal pitch processing (see Griffiths and Hall 2012), suggesting that such regions process pitch information regardless of its source. On the other hand, behavioral studies suggest that people with autism spectrum disorder are impaired in processing vocal pitch, but not pitch from other sources, implicating a specialized processing mechanism for vocal pitch (Jiang et al. 2015; Schelinski et al. 2017). More data are needed to determine whether there are specific vocal pitch-processing regions in the brain. Moreover, a brain region might be voice sensitive/selective only under specific circumstances. For example, several studies have found that the fusiform face area (FFA), a visual region specialized for face processing, responds more strongly when listeners try to recognize the identity of speakers whose faces they know than speakers whose faces are unknown (Sect. 7.3.3). This means that the FFA could be considered a voice-sensitive region, but only for a specific context (speaker recognition) and for specific stimuli (voices of speakers whose face is known).

7.3.2 Core Voice System

In their widely cited review of face processing, Haxby et al. (2000) proposed a distinction between a *core face system*, comprising brain regions directly involved in the visual analysis of faces, and an *extended face system*, comprising brain regions from nonvisual areas recruited to extract meaning from faces or to associate faces with stored information. Researchers have applied this idea to voice perception as well (Roswandowitz et al. 2017). Specifically, the core voice system comprises regions involved in the auditory analysis of voices, whereas the extended voice system comprises nonauditory regions. Here, we restrict our discussion to the brain regions most widely considered to belong to the core voice system, namely the posterior, anterior, and middle portions of the STS/STG, as well as the inferior frontal gyrus (IFG), some portion of which potentially belongs to the core voice system.

7.3.2.1 Posterior Superior Temporal Sulcus/Gyrus

The results of two fMRI studies suggest that the posterior STS/STG (pSTS/STG) plays a role in the analysis of VTL. In both of these studies, pSTS/STG bilaterally responded more strongly when listeners heard sequences of syllables synthesized to have varying VTL than when they heard sequences of syllables synthesized to have the same VTL (von Kriegstein et al. 2007, 2010). Note, however, that a third study with a similar design found more anterior STS/STG responses (von Kriegstein et al. 2006). In the von Kriegstein et al. (2007) study, listeners also heard sequences of a musical instrument and bullfrog vocalizations synthesized to have varying or the same spectral envelope (i.e., vocal tract length in bullfrogs and the equivalent in the musical instrument). The authors found that pSTS/STG bilaterally responded more strongly to spectral envelope variation only when listening to syllables, suggesting that it responded to human VTL variation specifically.

The pSTS/STG may play a role in the recognition of unfamiliar or recently familiarized speakers (Sect. 7.2.2.2). In a study by von Kriegstein and Giraud (2004), listeners completed two different tasks while listening to sequences of sentences spoken by different speakers. In one task (the speaker task), listeners heard a target sentence at the beginning of the sequence and judged whether sentences within the sequence were spoken by the target speaker. In another task (the speech task), listeners judged whether sentences within the sequence had the same linguistic content as the target sentence, regardless of the speaker. Within sequences, the sentences were either spoken by speakers who were all familiar (work colleagues) or by speakers who were all unfamiliar. The right pSTS/STG responded more strongly when listeners performed the speaker task than the speech task while listening to unfamiliar speakers; this response difference was stronger for the unfamiliar than for the familiar speakers (i.e., there was a task-by-familiarity interaction). It should be noted that this study was performed in the early days of fMRI research and some of its design features might be considered suboptimal by today's standards (e.g., the sample size was relatively small, and the data were analyzed using a fixed-effects model; see Talavage et al. 2012); eventually, these results must be replicated.

Zäske et al. (2017) provided supporting evidence for a role of the pSTS/STG in unfamiliar or recently familiarized speaker recognition. During an initial training phase, listeners became familiar with a set of speakers of German sentences (which were not intelligible, since none of the listeners spoke German). During MRI scanning, listeners heard the same or new sentences that were spoken by the same or new speakers. Listeners reported whether the speaker of each sentence was old or new, regardless of the sentence. The authors found stronger responses in several areas, including the right pSTS/STG, on trials when the listener correctly responded “new” than on trials when the listener correctly responded “old,” regardless of whether the sentence was new or old (i.e., a main effect of unfamiliarity). A possible limitation of this study is that listeners performed poorly on the task and were strongly biased toward responding “old” whenever they heard an old *sentence*, regardless of whether the *speaker* was new or old.

The precise computational role of the pSTS/STG during voice-identity recognition is not entirely clear. The results of the studies discussed above suggest that this region analyzes VTL, or serves a particularly important role for the recognition of unfamiliar speaker’s voices, or both. A current working hypothesis is that the pSTS/STG serves as a hub for establishing *reference patterns* for specific voice identities later committed to long-term memory (see Sect. 7.3.4.3) (Maguinness et al. 2018). This could explain why pSTS/STG particularly responds when attempting to recognize unfamiliar or recently familiarized voices because these are precisely the circumstances when reference patterns are required. Presumably, reference patterns are already established for highly familiar speakers.

According to Pernet et al. (2015), the posterior peak is the strongest and most consistent TVA elicited during passive listening to vocal sounds of unfamiliar speakers. However, it is not known whether the posterior TVA is the same as, incorporates, or is separate from the pSTS/STG region or regions implicated in the aforementioned studies, especially because the peak of the posterior TVA is highly variable across listeners.

7.3.2.2 Anterior Superior Temporal Sulcus/Gyrus

Studies suggest that the anterior STS/STG (aSTS/STG) plays a role in the representation of unique voice identities. In a study by von Kriegstein et al. (2003), listeners heard sequences of sentences and performed in alternating blocks a speaker task and a speech task (similar design as in von Kriegstein and Giraud 2004). The authors found that the right aSTS/STG responded more strongly during the speaker task than the speech task. This result must have been task related because listeners heard exactly the same stimuli during both tasks. The finding that the right aSTS/STG plays a role in voice-identity recognition was corroborated by a meta-analysis of relevant fMRI studies (Blank et al. 2014). Moreover, in a magnetoencephalography (MEG) study, Schall et al. (2014) found a positive correlation between behavioral performance and aSTS/G activity during a voice-identity recognition task. This region also appears to adapt to the presentation of different speech sounds from the same speaker (Belin and Zatorre 2003). Finally, in a study by Formisano et al. (2008), listeners repeatedly heard three vowels spoken by three speakers. When the authors applied multivariate pattern analysis to the fMRI data, the greatest concentration of informative voxels for classifying voice identity was found in the right aSTS/STG.

If the aSTS/STG represents unique voice identities, one might predict stronger aSTS/STG responses to familiar speakers than unfamiliar speakers. However, this does not appear to be the case; instead, familiar speakers elicit stronger responses within the extended voice system (e.g., von Kriegstein and Giraud 2004; von Kriegstein et al. 2005).

7.3.2.3 Middle Superior Temporal Sulcus/Gyrus

The role of the middle STS/STG (mSTS/STG) in voice processing and voice-identity recognition is unclear at present. A recent review article (Maguinness et al. 2018) proposed that the mSTS/STG plays a facilitative role by connecting the pSTS/STG, which performs the perceptual analysis required to perceive voice identity, with the aSTS/STG, which represents unique voice identities. Support for this hypothesis comes from the fMRI study by von Kriegstein and Giraud (2004), who examined functional connectivity (psychophysiological interactions) of brain regions during voice processing. The authors observed stronger connectivity between the right pSTS/STG and mSTS/STG and between the right aSTS/STG and mSTS/STG while listeners performed the speaker task than the speech task. Furthermore, using probabilistic tractography on diffusion-weighted imaging data, Blank et al. (2011) showed that all three TVAs are structurally connected with one another. One mysterious feature of the mSTS/STG is that it appears to be sensitive to voices (e.g., Belin et al. 2000; Pernet et al. 2015), yet it does not always appear to respond more strongly to voices during a speaker task than a speech task (e.g., Kriegstein and Giraud 2004).

7.3.2.4 Inferior Frontal Gyrus

Bilateral IFG was one of the extra-temporal regions implicated in voice-identity recognition in the meta-analysis by Blank et al. (2014). The authors performed separate meta-analyses of fMRI studies of voice identity, face identity, and name recognition, and they identified bilateral IFG clusters for both voice-identity and face-identity recognition. While these clusters were very close together, a conjunction analysis suggested that they were not the same, implying that the IFG is not involved in supra-modal person recognition. In their large-scale study, Pernet et al. (2015) also found voice sensitive maxima within left and right IFG whose locations seemed to be similar to those reported by Blank et al. (2014).

Whether the IFG should be considered part of the core or extended voice system is debatable. On the one hand, the meta-analysis by Blank et al. (2014) suggested that this region plays a critical role in auditory-only voice-identity recognition. However, the IFG also subsumes numerous other related processes: for example, the IFG famously includes Broca's area and is therefore critical for speech comprehension and production (see Hickok et al. 2011). The IFG is also part of the cortical pathway for discriminating vocal emotions (see Frühholz et al. 2016). In primates, ventrolateral prefrontal cortex (VLPFC), an area considered homologous with human IFG, responds to complex sounds, including species-specific vocalizations (Romanski and Goldman-Rakic 2002). However, single neurons in the VLPFC also respond to faces (O'Scalaidhe et al. 1997) and appear to integrate voice and face information (Sugihara et al. 2006).

7.3.3 Interactions Between Brain Regions

The following discussion of interactions between brain regions during voice processing is restricted to interactions between the core voice system and the FFA. The FFA is the most intensively studied face-sensitive region within the core face system that appears to serve a variety of functions related to face processing and whose precise role is a topic of ongoing research (for a review, see Kanwisher and Yovel 2006). The present chapter focuses on interactions of voice processing with this region specifically because they appear to be the most robust. Blank et al. (2014) provide a discussion of interactions that incorporate other regions.

Under normal circumstances, we are presented with voices and faces concurrently. Thus, if we are familiar with a speaker's voice, it is very likely that we are also familiar with his/her face. Over several studies, von Kriegstein and colleagues demonstrated that recognizing the voices of familiar speakers involves the FFA when both kinds of personal information are known. First, von Kriegstein et al. (2005) contrasted performance on speaker and speech tasks using stimuli from personally familiar speakers (whose faces were known to the listeners) and unfamiliar speakers (same data as in von Kriegstein and Giraud 2004). Listeners additionally completed a separate MRI scan to localize the FFA. The authors found that the bilateral fusiform gyrus responded specifically to recognizing the voices of familiar speakers. Importantly, this region overlapped with the FFA (defined using the separate face-area localizer scan), suggesting that familiar-speaker recognition indeed recruited the FFA.

A series of follow-up studies revealed a number of interesting features of the FFA during voice-identity processing: (1) the FFA was involved in auditory-only voice-identity recognition after only 2 min of audio-visual experience with the speaker; (2) the FFA selectively responded during speaker *recognition* and not, for example, during a speech task; (3) FFA responses were behaviorally relevant for voice-identity recognition—strength of FFA responses correlated positively with a measure of recognition performance (the face benefit) in neurotypical listeners and was reduced in individuals with face-identity recognition deficits; and (4) the FFA exhibited functional connectivity with the temporal voice-sensitive regions during voice-identity recognition, suggesting that these regions communicate with one another to resolve a speaker's identity (see von Kriegstein 2011; Mathias and von

Kriegstein 2014). Regarding point (4), functional connectivity between voice and face regions does not necessarily mean that those regions communicate directly; instead, it could reflect co-modulation by a later supra-modal region. Two studies provided evidence against this explanation. First, using probabilistic tractography, Blank et al. (2011) found direct white matter connections between the TVAs and the FFA (localized using fMRI). Connections were stronger between the middle/anterior TVAs and the FFA than between the posterior TVA and the FFA. Second, using MEG, Schall et al. (2013) found that the FFA activity during voice-identity recognition begins at an early stage of sensory processing, approximately 110 ms after sound onset. Those findings suggested that the functional correlations observed between the TVAs and FFA during speaker recognition are the result of direct communication rather than indirect co-modulation via a third region.

In summary, there is clear evidence of direct and early interactions between the TVAs and the FFA. These direct interactions serve to exchange physical and identity information about voices and faces (Blank et al. 2014). An open question is what specific computations are subserved by these cross-modal interactions. One hypothesis is that the brain simulates a talking face while we hear someone speaking and that this simulation process fills in the missing information normally available during audio-visual communication. Such simulation may help to better recognize both voice identity and the speech message (see von Kriegstein et al. 2008).

7.3.4 Deficits in Voice Processing

Our understanding of the neurobiology of voice-identity recognition has been improved by accounts of individuals with voice-processing disorders, particularly *phonagnosia.* Phonagnosia (from the Greek *phone*, "voice," and *agnosis*, "without knowledge") refers to a difficulty in recognizing voice identities. Van Lancker and Canter (1982) introduced the term to describe the auditory equivalent of prosopagnosia, a selective difficulty in recognizing faces (cf. Bodamer 1947). Phonagnosia may manifest following brain damage (*acquired phonagnosia*) or in the absence of brain damage, in which case it is presumably present at birth (*congenital* or *developmental phonagnosia*) (Garrido et al. 2009; Roswandowitz et al. 2014). It is also useful to draw a distinction between two potential forms of phonagnosia: a deficit in discriminating the features of voices, which leads to problems recognizing speakers (*apperceptive* phonagnosia) and a deficit in identifying familiar voices without discrimination problems (*associative* phonagnosia) (Hailstone et al. 2011; Roswandowitz et al. 2014). Readers interested in a complete account of the literature on phonagnosia are encouraged to consult the meta-analysis by Blank et al. (2014) and a recent article by Roswandowitz et al. (2018b), which provides an exhaustive, chronologically organized review of phonagnosia studies. Here, we focus our discussion on how the findings from such studies compare to those of functional neuroimaging studies involving healthy individuals.

7.3.4.1 Laterality

Early studies of acquired phonagnosia were concerned with answering the basic question of whether the left or right hemisphere of the brain is more important for voice processing. Although based on small numbers of cases (e.g., Van Lancker and Canter 1982), they invariably noted that, on average, patients with right-lateralized lesions performed worse than patients with left-lateralized lesions (or the control subjects) on tests of voice discrimination or recognition. Van Lancker et al. (1989) drew a more precise conclusion regarding the role of laterality in phonagnosia. They found that, on average, patients with unilateral brain lesions (either left or right) performed worse than controls on a test of unfamiliar-voice *discrimination*, in which they heard pairs of sentences and judged whether they were spoken by the same or different speakers. Importantly, on average, only the group with right-lateralized lesions performed worse than controls on a test of famous-voice *recognition*, in which they matched sentences spoken by famous people to their facial photographs and written names. In other words, lesions to either hemisphere caused apperceptive phonagnosia, whereas only right-hemisphere lesions caused associative phonagnosia.

The discussion of voice processing in the healthy human brain (Sects. 7.3.1–7.3.3) deliberately evaded the question of laterality. While several of the studies considered in those sections noted stronger or more consistent voice-related responses in the right hemisphere (e.g., Belin et al. 2000; von Kriegstein et al. 2003), they usually did not perform formal tests of lateralization. One study that did perform such tests found no related effects (Pernet et al. 2015). Note that the study by Pernet and colleagues involved passive listening to all kinds of vocal sounds (including speech) and, therefore, their results are not contradictory to Van Lancker et al. (1989), who had hypothesized that right-hemisphere lesions cause associative phonagnosia. The critical test is whether voice-identity *recognition* is right-lateralized in the healthy human brain. The results of a recent voxel-based lesion-behavior mapping study (Roswandowitz et al. 2018a), which is described in more detail later, were indeed consistent with this hypothesis. Specifically, patients with right-hemispheric lesions performed worse on a test of voice-identity recognition with recently familiarized voices, on average, than patients with left-hemispheric lesions.

7.3.4.2 Parietal Lobe Lesions

The role of the parietal lobe is a possible source of disagreement between functional neuroimaging studies involving healthy listeners and studies of acquired phonagnosia. No regions within the parietal lobe appear to be voice sensitive (Pernet et al. 2015), yet almost all reported cases of acquired phonagnosia have had parietal-lobe lesions (reviewed by Roswandowitz et al. 2018b). In an attempt to resolve this discrepancy, Roswandowitz et al. (2018a) performed voxel-based lesionbehavior mapping on a cohort of fifty-eight patients with focal unilateral brain lesions using a battery of tests. The battery included a test on familiar voice-identity recognition and two tests involved recently familiarized voices. In one of these tests, listeners heard sentences spoken by unfamiliar speakers presented concurrently with a written name, and the participants later matched these names to novel sentences from the same speakers (the voice-name test). The other test had the same design, but listeners matched speakers with faces (the voice-face test). The main result was that there was a relationship between lesions in the right temporal lobe and the right inferior parietal lobe and poor performance on all three tests of voice-identity recognition. Interestingly, lesions in the inferior parietal lobe (IPL) led to poor performance on the voice-face test but not the voice-name test. This striking partial dissociation suggested that the IPL plays a specific role in associating voices and faces.

Roswandowitz et al.'s (2018a) findings concerning the IPL were consistent with previous case studies of phonagnosia because the tests of famous-voice recognition employed in those studies involved matching voices to photographs of famous people (e.g., Van Lancker and Canter 1982; Van Lancker et al. 1989). Moreover, many previous functional neuroimaging studies of voice processing in healthy listeners did not involve faces (e.g., von Kriegstein et al. 2003; Pernet et al. 2015); so, from that perspective, it is not surprising that they did not implicate the IPL. In fact, two fMRI studies that did involve person-related voice and face information both found IPL responses. First, a study by von Kriegstein and Giraud (2006) found stronger responses to voice-identity recognition in the IPL after voice-face learning than voice-name learning. Note, however, that the functional connectivity between parietal regions and voice and face areas was stronger *before*, not after, learning voice-face associations. Second, Hölig et al. (2017) presented listeners with silent videos of faces speaking nonwords, followed by auditory-only presentations of the same words, and they found stronger responses of the IPL when the face and speaker were different than when they were the same person. Precisely what computational role the IPL plays in associating voices and faces is unclear at present, particularly in light of the evidence that voice and face regions are directly connected (Sect. 7.3.3).

7.3.4.3 Temporal Lobe Lesions

In the study by Roswandowitz et al. (2018a), temporal lobe lesions were associated with poor performance in all three voice-identity recognition tests (including famous and recently familiarized voices). This finding is consistent with the many neuroimaging studies implicating the TVAs during voice processing and voice-identity recognition (Sects. 7.3.1–7.3.3). In addition, the lesion-behavior associations in the posterior temporal lobe were stronger for the tests using recently familiarized voices than for the one using famous voices (Roswandowitz et al. 2018a). This result is also consistent with the finding that the pSTS/STG is particularly involved in processing relatively unfamiliar voices (Sect. 7.3.2.1).

One finding from Roswandowitz et al. (2018a) that is difficult to reconcile with previous neuroimaging studies is that lesions in the *anterior* temporal lobe were not associated with poor performance on the voice-identity recognition tests. Functional neuroimaging studies suggested that the pSTS/STG is involved in the auditory analysis of voices (Sect. 7.3.2.1), whereas the aSTS/STG plays a role in representing voice identity (Sect. 7.3.2.2). Therefore, one might expect strong relationships between voice-identity recognition and lesions in both of these regions.

Other studies have reported that lesions in the anterior temporal lobe are associated with deficits in supra-modal person recognition (e.g., Hailstone et al. 2011). Recently, however, Luzzi et al. (2018) reported a single case of seemingly pure associative phonagnosia for an individual who had a lesion in his anterior temporal lobe (also including lenticular and caudate nuclei). This patient found it difficult to identify singers from recordings he knew well. He also performed poorly on tests of familiar-voice recognition (personal acquaintances) and famous-voice recognition, yet performed normally on auditory, visual (including face-identity recognition), and several other control tasks. He was also unimpaired on unfamiliar voice discrimination (i.e., judging whether two auditory speech stimuli were spoken by the same or a different person). This last result is interesting because a similar dissociation between unfamiliar voice discrimination and familiar voice-identity recognition was reported in previous case studies (e.g., Van Lancker et al. 1989; discussed in detail by Maguinness et al. 2018). It is currently unclear whether this potential partial dissociation between unfamiliar-voice and familiar-voice processing and recognition is related to the different task demands or the different levels of familiarity with the voices (also see Maguinness et al. 2018).

7.3.4.4 Congenital Phonagnosia

To date, there have been five reported cases of voice-identity recognition deficits in the absence of brain damage. First, Garrido et al. (2009) found that KH, a 60 year-old female, performed worse than controls on tasks involving recognizing famous voices, learning and recognizing unfamiliar voices, and unfamiliar-voice discrimination, but she performed normally on tasks involving environmental-sound recognition, vocal affect perception, music perception, face recognition, as well as basic auditory and neuropsychological measures. Her phonagnosia was presumably apperceptive since impairments included both recognition and discrimination. KH had some difficulties with speech-in-noise perception, which the authors attributed to fatigue, but this also could have been because knowledge of voice features might help speech-in-noise comprehension (von Kriegstein et al. 2010).

Roswandowitz et al. (2014) reported two more cases of congenital phonagnosia: (1) AS, a 32 year-old female, and (2) SP, a 32 year-old male, both of whom performed worse than controls on tests of unfamiliar-voice learning and famous-voice recognition, but normally on auditory and visual control tasks. AS performed poorly on a voice-discrimination test, suggesting that she had apperceptive phonagnosia; SP performed normally on this task, suggesting that he had associative phonagnosia.

Xu et al. (2015) reported AN, a 20 year-old female who performed poorly on a test of famous-voice discrimination. This individual performed normally on a voice-discrimination test, suggesting that she had associative phonagnosia. Finally, Xu and colleagues came across SR, a 40 year-old male, but testing of this individual was not extensive. Whether SR was phonagnosic and what type of phonagnosia he had remain to be clarified.

The neural correlates of voice processing in AN were assessed via two fMRI experiments (Xu et al. 2015). First, the authors measured AN's brain activity during a voice-area localizer scan (cf., Belin et al. 2000), which revealed normal recruitment of the TVAs compared to controls. In the second experiment, AN and the controls were instructed to imagine a series of famous voices and nonvoice sounds. In AN, the contrast between voice and nonvoice imagery revealed reduced responses in the ventromedial prefrontal cortex, left precuneus, and left cuneus relative to controls. Whether dysfunction of these areas was to blame for AN's phonagnosia was unclear particularly because voice imagery and voice recognition are very different processes, and voice-imagery is likely a much more difficult task for AN than neurotypical controls (for further discussion, see Roswandowitz et al. 2017).

Roswandowitz et al. (2017) performed two fMRI experiments on AS and SP: a voice-area localizer scan (similar to Belin et al. 2000) and a scan contrasting a speaker task and a speech task performed using the same stimuli (same design as von Kriegstein et al. 2003). In both experiments, AS showed reduced responses in the core voice system, including the HG, PT, and the pTVA, compared to controls. This pattern was consistent with her symptoms: she experienced apperceptive phonagnosia, suggesting that she was impaired in the auditory analysis of voices. By contrast, SP, who experienced associated phonagnosia, exhibited typical responses of these areas but decreased connectivity between core and extended voice regions.

7.4 Models of Voice Processing and Voice-Identity Recognition

Contemporary models of voice processing come in two "flavors". First, *functional* or "box-and-arrow" models attempt to delineate the processes that logically must be applied to vocal sounds in order to gain meaningful information from them, typically the speech message, the emotional state of the speaker, and the speaker's identity (Sect. 7.4.1). While such models are obviously considerable simplifications of the real processes involved, they have nevertheless proved to be extremely useful for understanding voice processing because their components can be individually evaluated and mapped to specific brain regions using neuroscientific evidence. Models of the second "flavor" are concerned with voice-identity recognition specifically and attempt to explain how the brain encodes individual voice identities at an abstract level (Sect. 7.4.2). Currently, all models of the second flavor share the basic idea that voice identities are encoded in terms of their deviations from a *prototype* or average voice. The findings of several psychophysical studies are consistent with the notion of prototype-based encoding, and preliminary neuroscientific evidence points to where in the brain the prototypes might be stored and used for voice-identity recognition. A recently proposed model attempts to combine insights from both functional and prototype models (Sect. 7.4.3).

7.4.1 Functional Models

Historically, psychologists and neuroscientists have been more interested in faces than voices; consequently, functional models of voice processing have drawn heavily from prior models of face perception. Arguably, the most influential functional model of face processing was proposed by Bruce and Young (1986). According to this model, after basic visual analysis, facial representations are sent to modules that perform the analysis of facial speech, emotion expression, and face identity. The face-identity module is further broken down into *face-recognition units* (FRUs) that code specific facial identities that are familiar to the viewer and supra-modal *person-identity nodes* (PINs) that receive input from the FRUs. Finally, these three modules (speech, emotion, and identity) feed into a loosely defined, higher-order cognitive system for semantic processing.

Bruce and Young's (1986) model was later expanded by Ellis et al. (1997) to incorporate voices. The model of Ellis and colleagues assumed a functional system for voice processing exactly mirroring that for faces—including *voice-recognition units* (VRUs)—which merges with the face-processing system at the level of the PINs. The core tenets of the model have appeared in various forms in numerous reviews and empirical studies of voice perception (Neuner and Schweinberger 2000; Belin et al. 2011). Figure 7.5 illustrates two such models.

Much of the neuroscientific evidence presented in Sect. 7.3 is broadly consistent with functional models of voice processing. For instance, the distinction between the core and extended systems, originally proposed for faces (Haxby et al. 2000) but more recently extended to voices (e.g., Roswandowitz et al. 2017), fits neatly within some models (bottom panel of Fig. 7.5) (Roswandowitz et al. 2018b), with basic auditory and auditory module-specific analyses performed by the core voice system and semantic processing performed by the extended voice system. However, some specific findings are difficult to reconcile with the models. For example, the models predict that auditory voice-identity analysis, presumably performed by the pSTS/STG, occurs before voice-identity recognition, presumably performed by the aSTS/STG. However, in a MEG study by Schall et al. (2015), peaks in activity within the pSTS/STG and aSTS/STG occurred at roughly the same time when listeners identified speakers to whom they had been familiarized prior to scanning. Simultaneous pSTS/STG and aSTS/STG activity, assuming that their functions have been described correctly, raises the possibility that there is some degree of parallel processing within these regions when listeners hear moderately familiar speakers. In

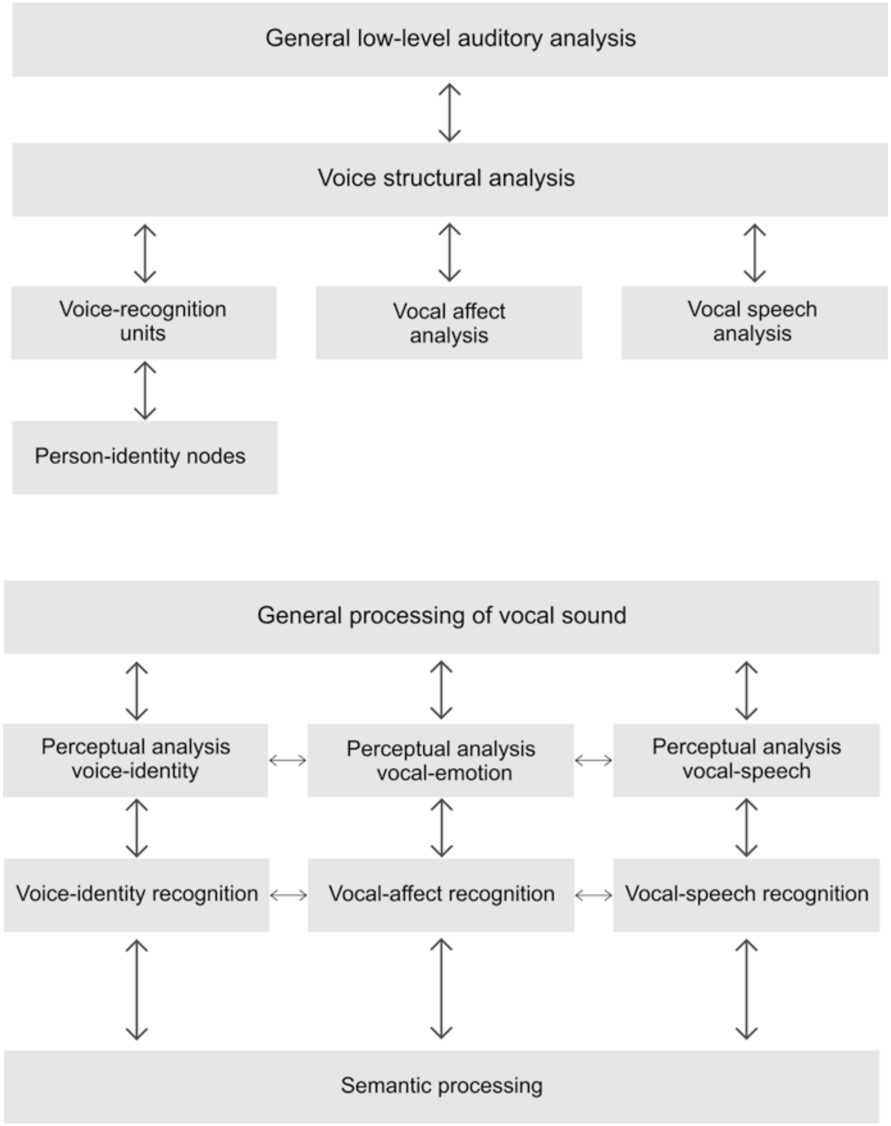
General low-level auditory analysis
Voice structural analysis
Voice-recognition units
Vocal affect analysis
Vocal speech analysis
Person-identity nodes
General processing of vocal sound
Perceptual analysis voice-identity
Perceptual analysis vocal-emotion
Perceptual analysis vocal-speech
Voice-identity recognition
Vocal-affect recognition
Vocal-speech recognition
Semantic processing

Fig. 7.5 Models of voice processing. **Top panel**: a schematic of the functional model of voice processing from Belin et al. (2011). **Bottom panel**: a schematic of another functional model with explicit interactions between auditory modules and a semantic-processing component. (Adapted from Maguinness et al. 2018)

addition, according to functional models of voice processing, there should be no divergence in unfamiliar-voice and familiar-voice processing until at least the level of the VRUs; however, neuroimaging and lesion reports suggest that there is already a dissociation between familiar-voice and unfamiliar-voice processing at the level of the proposed auditory analysis in pSTG/STS.

A further open question regarding functional models of voice processing is how best to characterize the interactions between the speech, emotion, and identity modules. Some models ignore such interactions completely (Belin et al. 2011) (top panel of Fig. 7.5). However, numerous psychophysical studies have shown that speech perception in noise is improved by familiarity with the speakers (Kreitewolf et al. 2017) and voice-identity recognition is much easier from speech in one's native language (Wester 2012), which should not be the case if speech and voice-identity information are processed completely independently. Other versions of the model do acknowledge these interactions (Roswandowitz et al. 2018b); however, to date, it is not clear exactly *how* and *when* these interactions occur.

Another gray area concerns the location and precise function of the PINs. So far there is only weak neuroscientific evidence for the existence of a dedicated module in the human brain that corresponds to the PINs (see Blank et al. 2014). Moreover, the fact that connections between the core voice system and the FFA are early (Schall et al. 2013) and direct (Blank et al. 2011) suggests that they are not necessarily comodulated by a third, supra-modal person-identity region. Taken to the extreme, these findings could be used to argue that PINs are simply not needed and that, instead, representations of supra-modal person-specific identities are fully encoded by modality-specific identity regions and the connections between them. This view is perhaps too extreme. Patients with frontotemporal dementia with right-hemispheric dominance often have impairment in supra-modal person-identity recognition (e.g., Gainotti et al. 2003), consistent with the hypothesis that their PINs have been eliminated. Then again, this finding could be explained by damage to distinct voice-processing and face-processing regions that are relatively close within the anterior temporal lobe. However, there is also fMRI evidence for the existence of PINs in the anterior temporal lobe of neurotypical brains (von Kriegstein and Giraud 2006).

7.4.2 Prototype Models

Prototype models have a very long history in psychology (Bartlett 1932) and have been used to explain encoding, categorization, and recognition of many kinds of stimuli. According to these models, the identity of a stimulus is encoded in terms of its deviations from an internal representation of a prototypical stimulus. In the context of voice-identity recognition, the prototype is an approximation of a typical voice built up by taking the average of our experiences with many different speakers. Figure 7.6 illustrates a prototype model of voice-identity recognition formulated by Lavner et al. (2001).

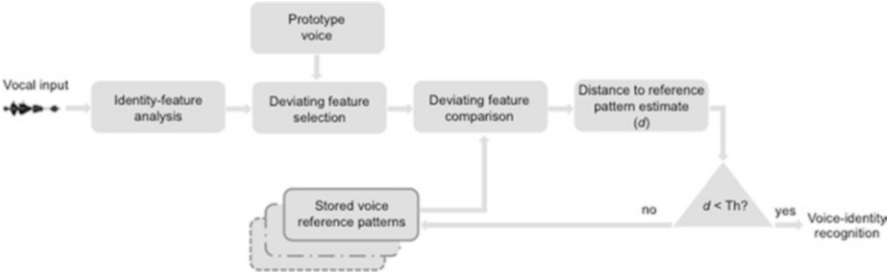
Prototype voice
Vocal input
Identity-feature analysis
Deviating feature selection
Deviating feature comparison
Distance to reference pattern estimate (*d*)
Stored voice reference patterns
no
d < Th?
yes
Voice-identity recognition

Fig. 7.6 Schematic of the prototype model by Lavner et al. (2001). According to this model, voice-identity features extracted from the stimulus are compared with those of a stored prototype. Features that deviate sufficiently from the prototype are selected and compared to stored reference patterns, one for each voice identity. Subtraction of the selected voice features from the closest reference pattern yields a distance measure, *d*. If *d* is smaller than some perceptual threshold, *Th*, the voice is deemed to belong to the identity corresponding to the reference pattern. If *d* exceeds the threshold, it is used to create a new reference pattern belonging to a new identity. (Adapted from Maguinness et al. 2018)

Findings from several behavioral studies are broadly consistent with general principles of prototype-based encoding of voice identity. For example, ear-witness reliability is subject to a typicality effect, such that speakers with unusual or distinctive voices are easier to remember than those with typical voices (Mullennix et al. 2011). Distinctive voices should be easier to remember than typical ones according to the prototype model because they have fewer “near neighbors” to compete with. Moreover, famous-speaker recognition appears to be subject to a caricature effect. López et al. (2013) instructed professional impersonators to imitate the voices of famous Spanish speakers based on their memory of the speakers and then again after listening to recordings of the speakers. Naïve listeners were better at recognizing the famous speakers from the first set of impersonations than the second set, whereas the second set was judged to be more similar to the original recordings than the first set. The prototype model predicts better recognition from the first set of impersonations because those impersonations emphasized the deviations from the prototype (i.e., were caricatures), which are more important for recognition than acoustic accuracy. Finally, perceptual aftereffects following adaptation to voices are also consistent with the prototype model (Latinus and Belin 2011).

To date, just one study has examined the neurobiology of prototype-based encoding of voice identity (Latinus et al. 2013). In this study, the authors recorded brain responses with fMRI while listeners heard recordings of many speakers (32 male and 32 female) speaking either the word “had” (in one experiment) or “hello” (in another experiment). From each recording within an experiment, they measured mean f_0 (i.e., GPR), mean distance between f_1 and f_2 (somewhat related to VTL), and harmonics-to-noise ratio (HNR). They also synthesized average male and female voices, and for each veridical recording, calculated the Euclidean distance in terms of the three acoustic features from their sex-specific (and word-specific) average.

The results suggested that the TVAs are sensitive to deviations from a prototypical voice (see also the discussion in Maguinness et al. 2018). However, this interpretation is complicated by the fact that correlations were strong when the authors assumed separate male and female averages, yet the correlations were weak when they assumed a single, androgynous average. The authors also constructed different prototypes for their two experiments. Thus, it appeared that listeners generated separate male and female prototypes and separate prototypes for different words.

A limitation of prototype models is that, to date, no alternatives have been proposed or tested, which makes it difficult to properly evaluate them. In other areas of cognitive psychology, prototype models are frequently contrasted with *exemplar* models (Nosofsky 1986). The subtle yet important distinction between prototype and exemplar models is that under the former, a stimulus is encoded relative to an internally constructed prototype, whereas under the latter, a stimulus is encoded relative to representations of previously encountered stimuli from the set. Prototype and exemplar models frequently yield similar predictions, but they could be differentiated in the future.

7.4.3 A New Model

Recently, a new model has been proposed that attempts to combine prior functional and prototype models and to incorporate the finding of a partial dissociation between unfamiliar-voicw and familiar-voice processing (Fig. 7.7) (Maguinness et al. 2018). Unlike some previous functional models (Sect. 7.4.1), this model focuses on voice-identity and face-identity recognition, although it could be extended to include speech and emotion processing in the visual and auditory modality in the future.

The new model assumes a first step of basic auditory analysis (*identity-feature analysis*). This step is obligatory and similar to the feature-extraction step from the prototype model of Lavner et al. (2001) (Sect. 7.4.2). It is different from (most) functional models (Sect. 7.4.1) because it assumes a certain level of specialization for voice-identity features; in previous functional models the earliest processing step is usually not considered to differ for voice-identity, speech, and vocal-emotion analyses.

Following identity-feature analysis, extracted vocal features are compared to those of a prototype (or several prototypes) in order to select deviant features. In the next step, deviant features are compared to stored reference patterns. Subtraction of the deviant voice features from the closest reference pattern yields a distance measure, d. If $d < Th$ (some perceptual threshold), the voice is deemed to belong to the identity corresponding to the closest reference pattern, and voice processing continues within the anterior STG/STS of the core-voice system and the extended voice system. If $d > Th$, a voice enters the *perceptual voice-identity processing loop* (light grey arrows in Fig. 7.7), which serves to establish new reference patterns. Reference-pattern establishment requires some information to be sent back for reanalysis of voice-identity features. Importantly, this loop is engaged while

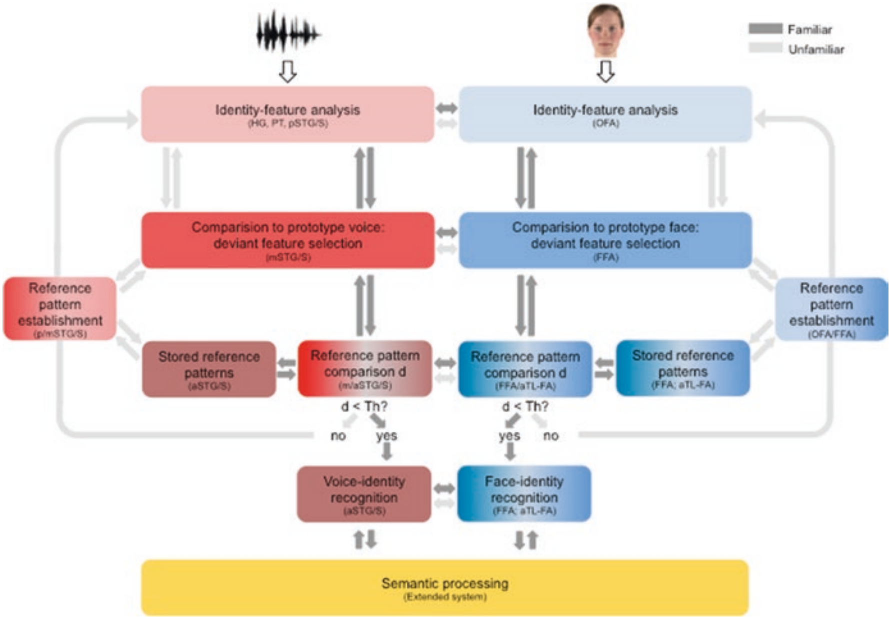

Fig. 7.7 The new model of voice-identity and face-identity recognition. The *colored boxes* represent the various processing steps in recognizing voices and faces, with *grey lines* representing the flow of information. Functions are matched to their assumed functions, although at the moment this matching should be considered tentative. *a*, anterior*; aTL-FA*, anterior temporal lobe/face area; *d*, difference between voice and reference pattern; *FFA*, fusiform face area; *HG*, Heschl's gyrus; *m*, middle; *OFA*, occipital face area; *p*, posterior; *PT*, planum temporale; *STS/G*, superior temporal sulcus/gyrus; *Th*, threshold. (Adapted from Maguinness et al. 2018)

listeners are becoming familiar with a speaker's voice. The loop may be engaged more strongly if listeners are trying to learn new speakers or trying to recognize recently familiarized speakers. Moreover, the loop may not be engaged for highly familiar speakers for whom strong reference patterns have already been established.

This proposal is consistent with the finding that part of the core voice system (specifically, the pSTS/STG) responds more strongly to unfamiliar speakers than familiar speakers during speaker recognition. It is also a potential explanation for why unfamiliar-voice and familiar-voice processing might at least partly dissociate: Lesions to particular regions involved in the loop will lead to stronger impairments in unfamiliar or recently familiarized voice-identity recognition than familiar voice-identity recognition. Consistent with prior functional models (Sect. 7.4.2), the new model assumes that voice-identity and face-identity recognition involve equivalent, parallel processes that are localized to different voice-sensitive and face-sensitive brain regions, respectively.

An important feature of the new model is that it acknowledges the existence of connections (direct or indirect, functional or structural) between the voice-processing and face-processing pathways at various stages. Given the assumed functions of brain regions at these stages, the model allows us to speculate about what the connections between them might accomplish. For example, according to the model, the direct functional connections between the mSTS/STG and the FFA (reported by Blank et al. 2011) suggest that information from one modality might influence the selection of deviant features when recognizing an individual from stimuli presented to the other modality (also see Blank et al. 2014).

7.5 Summary

The human voice is the most important sound source in our environment, not only because it produces speech, but also because it conveys a wealth of information about the speaker. In many situations, listeners are able to simultaneously understand the speech message and identify the speaker with minimal effort. Psychophysical studies have investigated what acoustic and perceptual features (i.e., voice qualities) distinguish voices from different speakers and which of these features listeners use to recognize who is speaking. Glottal and vocal tract characteristics strongly influence perceived similarity between speakers and may serve as cues for voice-identity recognition. However, the importance of a particular feature strongly depends on the speaker and the stimulus.

Voice-identity recognition relies on a network of brain regions comprising a core voice system of several auditory regions within the temporal lobe, including regions dedicated to processing glottal and vocal tract characteristics and regions that play more abstract roles, and an extended voice system of non-auditory regions involved in the retrieval of information associated with specific voice identities (e.g., faces and names). Surprisingly, this network is supported by early, direct connections between regions within the core voice system and an analogous core face system, which serve to optimize voice-identity recognition.

Despite considerable recent advances in our understanding of human voice processing and voice-identity recognition, many questions remain unanswered. For example, when and where does the processing of unfamiliar and familiar speakers' voices diverge in the brain? Several functional neuroimaging studies have investigated differences in responses to speech from unfamiliar and familiar speakers, but we are far from a complete understanding of this issue. A related question is when and where voice discrimination and recognition dissociate. Unfortunately, both factors are usually intertwined: tests of voice discrimination tend to use speech from unfamiliar speakers; tests of recognition tend to use speech from familiar speakers. A potential way forward is to study more individuals with brain lesions and/or individuals with specific voice-processing deficits (e.g., phonagnosia) and use more sophisticated tasks and stimuli in comprehensive behavioral test batteries.

Future studies could work to improve theoretical models of voice processing. Prototype models of voice-identity encoding, in particular, are currently quite vague. Where are prototypes stored in the brain and which regions are responsible for generating them? How many prototypes are there? Could voice-identity encoding be more accurately described as exemplar based?

To date, the fine-mapping of brain regions to their functions during voice processing remains tentative. We still do not know why, for example, the aSTS/STG is not particularly sensitive to familiar-voice identities, whether the IFG should be considered part of both the core voice and core face systems, or whether the IPL is truly important for matching faces to voices, as suggested by lesion studies. Answers to these questions are likely to come from combinations of psychophysical and various neuroimaging techniques in neurotypical participants as well as those with voice-identity recognition deficits.

Compliance with Ethics Statements Samuel Robert Mathias declares that he has no conflict of interest.

Katharina von Kriegstein declares that she has no conflict of interest.

References

Agus TR, Paquette S, Suied C et al (2017) Voice selectivity in the temporal voice area despite matched low-level acoustic cues. Sci Rep 7(1):11526

Andics A, Gácsi M, Faragó T et al (2014) Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. Curr Biol 24(5):574–578

Baumann O, Belin P (2010) Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychol Res 74(1):110–120

Bartlett FC (1932) Remembering: a study in experimental and social psychology. Cambridge University Press, Cambridge

Belin P, Bestelmeyer PEG, Latinus M, Watson R (2011) Understanding voice perception. Br J Psychol 102(4):711–725

Belin P, Zatorre RJ, Ahad P (2002) Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res 13(1):17–26

Belin P, Zatorre RJ (2003) Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14(16):2105–2109

Belin P, Zatorre RJ, Lafaille P et al (2000) Voice-sensitive areas in human auditory cortex. Nature 403(6767):309–312

Blank H, Anwander A, von Kriegstein K (2011) Direct structural connections between voice- and face-recognition areas. J Neurosci 31(36):12906–12915

Blank H, Wieland N, von Kriegstein K (2014) Person recognition and the brain: merging evidence from patients and healthy individuals. Neurosci Biobehav Rev 47:717–734

Bodamer J (1947) Die Prosop-Agnosie (Prosopagnosia) Archiv für Psychiatrie und Nervenkrankheiten (Archive for Psychiatry and Neurological Diseases) 179(1–2):6–53

Bruce V, Young A (1986) Understanding face recognition. Br J Psychol 77(3):305–327

Ellis H, Jones D, Mosdell N (1997) Intra- and inter-modal repetition priming of familiar faces and voices. Br J Psychol 88(1):143–156

Fecteau S, Armony JL, Joanette Y, Belin P (2004) Is voice processing species-specific in human auditory cortex? An fMRI study. NeuroImage 23(3):840–848

Fitch WT, Giedd J (1999) Morphology and development of the human vocal tract: a study using magnetic resonance imaging. J Acoust Soc Am 106(3):1511–1522

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322(5903):970–973

Fouquet M, Pisanski K, Mathevon N, Reby D (2016) Seven and up: individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood. R Soc Open Sci. https://doi.org/10.1098/rsos.160395

Frühholz S, Trost W, Kotz SA (2016) The sound of emotions — Towards a unifying neural network perspective of affective sound processing. Neurosci Biobehav Rev 68:96–110

Gainotti G, Barbier A, Marra C (2003) Slowly progressive defect in recognition of familiar people in a patient with right anterior temporal atrophy. Brain 126(4):792–803

Garrido L, Eisner F, McGettigan C et al (2009) Developmental phonagnosia: a sensitive deficit of vocal identity recognition. Neuropsychologia 47:123–131

Gaudrain E, Li S, Ban V, Patterson RD (2009) The role of glottal pulse rate and vocal tract length in the perception of speaker identity. Paper presented at Interspeech 2009: 10th annual conference of the international speech communication association, 1–5, 148–151

Gilbert HR, Weismer GG (1974) The effects of smoking on the speaking fundamental frequency of adult women. J Psycholinguist Res 3(3):225–231

Gray H (1918) Anatomy of the human body. Lea Febiger, Philadelphia

Griffiths TD, Hall DA (2012) Mapping pitch representation in neural ensembles with fMRI. J Neurosci 32(39):13343–13347

Hailstone JC, Ridgway GR, Bartlett JW et al (2011) Voice processing in dementia: a neuropsychological and neuroanatomical analysis. Brain 134:2535–2547

Hautamäki R, Kinnunen T, Hautamäki V, Laukkanen A-M (2015) Automatic versus human speaker verification: the case of voice mimicry. Speech Comm 72:13–31

Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. Trends Cogn Sci 4(6):223–233

Hickok G, Costanzo M, Capasso R, Miceli G (2011) The role of Broca's area in speech perception: evidence from aphasia revisited. Brain Lang 119(3):214–220

Hillenbrand J, Getty LA, Clark MJ, Wheeler K (1995) Acoustic characteristics of American English vowels. J Acoust Soc Am 97(5):3099–3111

Hillenbrand JM, Clark MJ (2009) The role of f0 and formant frequencies in distinguishing the voices of men and women. Atten Percept Psychophys 71(5):1150–1166

Hölig C, Föcker J, Best A et al (2017) Activation in the angular gyrus and in the pSTS is modulated by face primes during voice recognition. Hum Brain Mapp 38(5):2553–2565

Hollien H, Shipp T (1972) Speaking fundamental frequency and chronologic age in males. J Speech Lang Hear Res 15(1):155–159

Jiang J, Liu F, Wan X, Jiang CM (2015) Perception of melodic contour and intonation in autism spectrum disorder: evidence from Mandarin speakers. J Autism Dev Disord 45:2067–2075

Johnson K (2005) Speaker normalization in speech perception. In: Pisoni DP, Remez RR (eds) The handbook of speech perception. Blackwell Publishing Ltd, Malden, pp 363–389

Kanwisher N, Yovel G (2006) The fusiform face area: a cortical region specialized for the perception of faces. Philos Trans R Soc Lond Ser B Biol Sci 361(1476):2109–2128

Kell AJ, Yamins DL, Shook EN et al (2018) A task-optimized neural network replicates human auditory behavior predicts brain responses and reveals a cortical processing hierarchy. Neuron 98:630–644

Kitaoka N, Enami D, Nakagawa S (2014) Effect of acoustic and linguistic contexts on human and machine speech recognition. Comput Speech Lang 28(3):769–787

Kreiman J, Vanlancker-Sidtis D, Gerratt BR (2005) Perception of voice quality. In: Pisoni DP, Remez RR (eds) The handbook of speech perception. Blackwell Publishing Ltd., Malden, pp 338–362

Kreiman J, Gerratt BR (1998) Validity of rating scale measures of voice quality. J Acoust Soc Am 104(3):1598–1608

Kreitewolf J, Gaudrain E, von Kriegstein K (2014) A neural mechanism for recognizing speech spoken by different speakers. NeuroImage 91:375–385

Kreitewolf J, Mathias SR, von Kriegstein K (2017) Implicit talker training improves comprehension of auditory speech in noise. Front Psychol. https://doi.org/10.3389/fpsyg.201701584

Künzel HJ (1989) How well does average fundamental frequency correlate with speaker height and weight? Phonetica 46(1–3):117–125

Latinus M, Belin P (2011) Anti-voice adaptation suggests prototype-based coding of voice identity. Front Psychol 2:175

Latinus M, McAleer P, Bestelmeyer PEG, Belin P (2013) Norm-based coding of voice identity in human auditory cortex. Curr Biol 23(12):1075–1080

Laver J (1980) The phonetic description of voice quality. Cambridge University Press, Cambridge

Lavner Y, Gath I, Rosenhouse J (2000) The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. Speech Comm 30:9–26

Lavner Y, Rosenhouse J, Gath I (2001) The prototype model in speaker identification by human listeners. Int J Speech Technol 4(1):63–74

López S, Riera P, Assaneo MF et al (2013) Vocal caricatures reveal signatures of speaker identity. Sci Rep. https://doi.org/10.1038/srep03407

Luzzi S, Coccia M, Polonara G et al (2018) Sensitive associative phonagnosia after right anterior temporal stroke. Neuropsychologia 116:154–161. https://doi.org/10.1016/j.neuropsychologia.2017.05.016

Maguinness C, Roswandowitz C, von Kriegstein K (2018) Understanding the mechanisms of familiar voice-identity recognition in the human brain. Neuropsychologia 166:179–193

Mathias SR, von Kriegstein K (2014) How do we recognise who is speaking. Front Biosci S6:92–109

Mullennix JW, Ross A, Smith C, Kuykendall K, Conrad J, Barb S (2011) Typicality effects on memory for voice: implications for earwitness testimony. Appl Cogn Psychol 25(1):29–34

Murray T, Singh S (1980) Multidimensional analysis of male and female voices. J Acoust Soc Am 68(5):1294–1300

Neuner F, Schweinberger SR (2000) Neuropsychological impairments in the recognition of faces voices and personal names. Brain Cogn 44(3):342–366

Nosofsky RM (1986) Choice similarity and the context theory of classification. J Exp Psychol Learn Mem Cogn 10:104–114

O'Scalaidhe SP, Wilson FA, Goldman-Rakic PS (1997) Areal segregation of face-processing neurons in prefrontal cortex. Science 278(5340):1135–1138

Petkov CI, Kayser C, Steudel T et al (2008) A voice region in the monkey brain. Nat Neurosci 11(3):367–374

Pernet CR, McAleer P, Latinus M et al (2015) The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. NeuroImage 119:164–174

Perrodin C, Kayser C, Logothetis NK, Petkov CI (2011) Voice cells in the primate temporal lobe. Curr Biol 21(16):1408–1415

Peterson GE, Barney HL (1952) Control methods used in a study of the vowels. J Acoust Soc Am 24(4):175–184

Plack CJ, Oxenham AJ (2005) The psychophysics of pitch. In: Plack CJ, Oxenham AJ, Popper AN, Fay RR (eds) Pitch: neural coding and perception. Springer Handbook of Auditory Research, vol 24. Springer, New York, pp 7–55

Remez RE, Fellowes JM, Rubin PE (1997) Talker identification based on phonetic information. J Exp Psychol Hum Percept Perform 23(3):651–666

Romanski LM, Goldman-Rakic PS (2002) An auditory domain in primate prefrontal cortex. Nat Neurosci 5(1):15–16

Roswandowitz C, Kappes C, Obrig H, von Kriegstein K (2018a) Obligatory and facultative brain regions for voice-identity recognition. Brain 141(1):234–247

Roswandowitz C, Maguinness C, von Kriegstein K (2018b) Deficits in voice-identity processing: acquired and developmental phonagnosia. In: Frühholz S, Belin P (eds) The oxford handbook of voice perception. Oxford University Press, Oxford

Roswandowitz C, Mathias SR, Hintz F et al (2014) Two cases of sensitive developmental voice-recognition impairments. Curr Biol 24(19):2348–2353

Roswandowitz C, Schelinski S, von Kriegstein K (2017) Developmental phonagnosia: linking neural mechanisms with the behavioural phenotype. NeuroImage 155:97–112

Saslove H, Yarmey AD (1980) Long-term auditory memory: Speaker identification. J Appl Psychol 65(1):111–116

Schall S, Kiebel SJ, Maess B, von Kriegstein K (2013) Early auditory sensory processing of voices is facilitated by visual mechanisms. NeuroImage 77:237–245

Schall S, Kiebel SJ, Maess B, von Kriegstein K (2014) Voice identity recognition: functional division of the right STS and its behavioral relevance. J Cogn Neurosci 27(2):280–291

Schall S, Kiebel SJ, Maess B, von Kriegstein K (2015) Voice identity recognition: functional division of the right STS and its behavioral relevance. J Cogn Neurosci 27(2):280–291

Schelinski S, Roswandowitz C, von Kriegstein K (2017) Voice identity processing in autism spectrum disorder. Autism Res 10(1):155–168

Sheffert SM, Pisoni DB, Fellowes JM, Remez RE (2002) Learning to recognize talkers from natural sinewave and reversed speech samples. J Exp Psychol Hum Percept Perform 28(6):1447–1469

Smith DRR, Patterson RD (2005) The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. J Acoust Soc Am 118(5):3177–3186

Smith DRR, Patterson RD, Turner R et al (2005) The processing and perception of size information in speech sounds. J Acoust Soc Am 117(1):305–318

Stevenage SV, Clarke G, McNeill A (2012) The "other-accent" effect in voice recognition. J Cogn Psychol 24(6):647–653

Stoicheff ML (1981) Speaking fundamental frequency characteristics of nonsmoking female adults. J Speech Lang Hear Res 24(3):437–441

Sugihara T, Diltz MD, Averbeck BB, Romanski LM (2006) Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. J Neurosci 26(43):11138–11147

Talavage TM, Johnsrude IS, Gonzalez-Castillo J (2012) In: Poeppel D, Overath T, Popper AN, Fay RR (eds) The human auditory cortex. Springer handbook of auditory research, vol 43. Springer, New York, pp 129–164

Titze I (1989) Physiologic and acoustic differences between male and female voices. J Acoust Soc Am 85(4):1699–1707

van Lancker D, Kreiman J, Emmorey K (1985) Familiar voice recognition: patterns and parameters. Part I Recognition of backward voices. J Phon 13:19–38

van Lancker DR, Canter GJ (1982) Impairment of voice and face recognition in patients with hemispheric damage. Brain Cogn 1:185–195

van Lancker DR, Kreiman J, Cummings J (1989) Voice perception deficits: neuroanatomical correlates of phonagnosia. J Clin Exp Neuropsychol 11(5):665–674

von Kriegstein K (2011) A multisensory perspective on human auditory communication. In: Murray MM, Wallace MT (eds) The neural bases of multisensory processes. CRC Press, Boca Raton, pp 683–700

von Kriegstein K, Dogan O, Grüter M et al (2008) Simulation of talking faces in the human brain improves auditory speech recognition. Proc Natl Acad Sci U S A 105(18):6747–6752

von Kriegstein K, Kleinschmidt A, Giraud A (2006) Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. Cereb Cortex 16(9):1314–1322

von Kriegstein K, Eger E, Kleinschmidt A, Giraud A-L (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. Cogn Brain Res 17(1):48–55

von Kriegstein K, Giraud A-L (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. NeuroImage 22(2):948–955

von Kriegstein K, Giraud A-L (2006) Implicit multisensory associations influence voice recognition. PLoS Biol 4(10). https://doi.org/10.1371/journal.pbio.0040326

von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud A-L (2005) Interaction of face and voice areas during speaker recognition. J Cogn Neurosci 17(3):367–376

von Kriegstein K, Kleinschmidt A, Giraud A (2006) Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. Cereb Cortex 16(9):1314–1322

von Kriegstein K, Smith DRR, Patterson RD et al (2007) Neural representation of auditory size in the human voice and in sounds from other resonant sources. Curr Biol 17(13):1123–1128 von Kriegstein K, Smith DRR, Patterson RD et al (2010) How the human brain recognizes speech in the context of changing speakers. J Neurosci 30(2):629–638

Wester M (2012) Talker discrimination across languages. Speech Comm 54:781–790

Wilding J, Cook S (2000) Sex differences and individual consistency in voice identification. Percept Mot Skills 91(2):535–538

Xu X, Biederman I, Shilowich BE et al (2015) Developmental phonagnosia: Neural correlates and a behavioral marker. Brain Lang 149:106–117

Yarmey AD (2007) The psychology of speaker identification and earwitness memory. In: Lindsay RCL, Ross DF, Read JD, Toglia MP (eds) The handbook of eyewitness psychology vol II: memory for people. Lawrence Erlbaum Associates, Mahwah, pp 101–136

Zäske R, Hasan BAS, Belin P (2017) It doesn't matter what you say: fMRI correlates of voice learning and recognition independent of speech content. Cortex 94:100–112

Chapter 8
Timbre as a Structuring Force in Music

Stephen McAdams

Abstract The study of timbre by music researchers is seriously underdeveloped in both the humanities and human sciences. As applied to instrumental music, theories to explain instrumental combinations and timbral shaping through instrumentation and orchestration are rare. Analyses of orchestration treatises and musical scores reveal an implicit understanding of auditory grouping principles by which many orchestral techniques and their perceptual effects function. This chapter, with a primary focus on classical Western orchestral and electroacoustic music, discusses connections between orchestration practice and perceptual principles based on research in auditory scene analysis and timbre perception. The chapter explores: (1) listeners' abilities to perceive relations among timbres; (2) how concurrent grouping cues result in blended or heterogeneous combinations of instruments; (3) how sequential groupings into segregated melodic streams and stratified foreground and background layers are influenced by timbral similarities and differences; and (4) how segmental grouping cues based on changes in instrument timbre and instrumental textures create musical units, formal boundaries, and expressive shaping of timbre melodies and larger-scale orchestral gestures.

Keywords Auditory grouping · Auditory scene analysis · Auditory stream segregation · Concurrent grouping · Orchestral gesture · Perceptual fusion · Segmental grouping · Sequential grouping · Timbre blend · Timbre contrast · Timbre interval

S. McAdams (✉)
Schulich School of Music, McGill University, Montreal, QC, Canada
e-mail: stephen.mcadams@mcgill.ca

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_8

8.1 Introduction

Timbre perception is at the heart of orchestration practice, that is, the choice, combination, and juxtaposition of instruments to create a specific musical effect. Examples include picking a particular instrument for the emotional tone it can convey, such as the melancholy English horn in the third act of Wagner's opera "Tristan und Isolde", or bouncing musical patterns between contrasting instrument families as in the second movement of Beethoven's ninth Symphony (where a repeating call and response pattern alternates between woodwinds plus brass and strings plus brass). For practicing composers and conductors, the potential of timbre to structure musical forms and to sculpt music's emotional impact is evident; however, until relatively recently (see Thoret et al. 2018), these roles of timbre have been addressed only rarely in music research in both the humanities (music theory, musicology, ethnomusicology) and the behavioral sciences (experimental psychology). Most researchers and theorists focus on the musical parameters of pitch and duration that give rise to melody and harmony on the one hand and rhythm and meter on the other (obviously in concert with other so-called "secondary" parameters such as loudness or musical dynamics and timbre).

An examination of writings on orchestration practice from the middle of the nineteenth century to present times (e.g., Berlioz and Strauss 1948; Adler 2002) reveals that the communication of knowledge about orchestration is primarily based on a multitude of examples of various techniques. From these, students must memorize all the cases or somehow implicitly derive theory by studying scores and listening carefully to recordings over a span of many years. They also learn by practicing orchestration techniques, with the added difficulty of not always being able to hear the musical result of what they might write on the page of a score because they rarely have access to an orchestra.

An alternative approach would be to start with the assumption that orchestration conventions have some implicit basis in auditory grouping principles, given that composers are most likely grounding what they do in their own auditory experience (Goodchild and McAdams 2018). As detailed in Sect. 8.3, auditory grouping includes the perceptual fusion of concurrent acoustic information into auditory events, the perceptual connection through time of similar events into auditory streams (melodic lines) or foreground and background layers, and the segmentation of streams or layers into "chunks" that can be processed in short-term memory. Timbre arises from perceptual fusion as a property of an event. It can then influence the way successive events form auditory streams because listeners tend to connect events coming from the same sound source and because, generally speaking, a given source varies relatively little in timbre compared to the differences between distinct sources. Timbral contrasts can provoke segmentation in which successions of events with similar timbres form units separated from preceding or succeeding material with different timbres. From this perspective, the role of timbre as a structuring force in music can be addressed through the following set of questions:

- Can relations among timbres in sequences be perceived, stored in memory, and subsequently recognized as intervals or contours analogous to the perception of pitch and duration relations?

- What is the relation between auditory fusion and the perception of timbre?
- In what way do auditory-scene-analysis principles and acoustic properties contribute to determining whether separate sound events will blend together?
- How do timbral continuity and discontinuity contribute to the formation of auditory streams and the formation of foreground and background orchestral layers?
- How do timbral changes and the learning of timbral patterns affect perceptual segmentation of sequences, and what role do these play in music?
- How do gradual and sudden timbral changes contribute to larger-scale musical form?

8.2 Perception of Timbral Relations

One of the properties of musical pitch that endows it with its psychological capacity to serve as a vehicle for musical patterns and forms is that *relations* among pitches (contours or intervals) can be perceived as musical qualities per se. Musical patterns can be constructed with these qualities, and operations on those patterns that maintain the structural relations, such as transposition, also maintain a strong degree of perceptual similarity between the original and transformed materials. For example, someone can hum the tune to *Happy Birthday* starting on any pitch and, if the intervals are correct, the melody is still recognized. In order to extend these form-bearing possibilities of pitch into the realm of timbre, it would be necessary to determine the kinds of structuring of timbral relations that can be perceived by listeners and that still provide a certain richness to be reasoned with by composers. For the psychologist, several interesting questions arise concerning a listener's ability to perceive and remember timbral relations in tone sequences and to build up hierarchical mental representations based on those relations (McAdams 1989).

Timbre space provides a model for relations among timbres. A *timbre space* is derived from dissimilarity ratings on all pairs of a set of sounds (usually equalized for pitch, duration, and loudness) to which a multidimensional scaling algorithm is applied to model the dissimilarities as distances in a Euclidean space (for more detail, see McAdams, Chap. 2). Sounds with similar timbres are close in the space and different ones are farther apart. The dimensions are presumed to be perceptual. A timbre interval can be considered as a vector connecting two timbres in such a space, and transposing that interval maintains the same amount of change along each perceptual dimension of timbre. One might ask whether listeners can perceive timbral intervals and recognize transpositions of those intervals to other points in the timbre space as one can perceive pitch intervals and their transpositions in pitch space. Consider the timbral trajectory shown in Fig. 8.1 through the McAdams et al. (1995) timbre space starting with the *guitarnet* (a synthetic hybrid of guitar and clarinet) and ending with the English horn imitation. How would one construct a timbre sequence starting from the bowed string so that it would be perceived as a transposition of this *Klangfarbenmelodie* (the German term for tone color melody

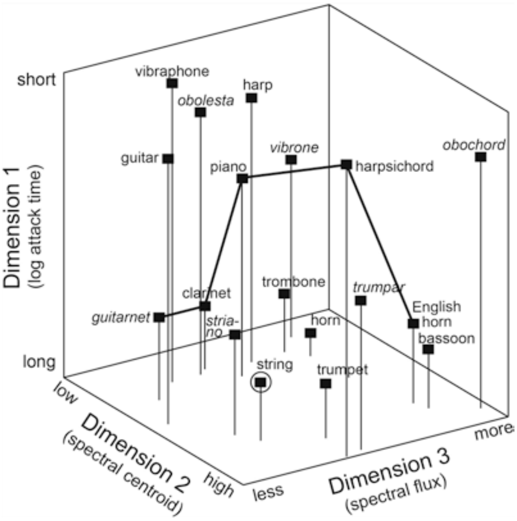

Fig. 8.1 A trajectory (*heavy black line*) of a short timbre melody through timbre space of synthesized sounds intended to mimic acoustical instruments or hybrids (*in italics*). How would one transpose the timbre melody that starts on *guitarnet* to a melody starting on string (*circled*)?

introduced by Schoenberg [1978])? If timbre interval perception can be demonstrated, the door would be opened for the application of some of the operations commonly used on pitch sequences to timbre sequences (Slawson 1985). The perceptual interest of this possibility is that it would extend the use of the timbre space as a perceptual model beyond the dissimilarity paradigm used to construct it in the first place.

Ehresman and Wessel (1978) first conceived of the notion of a timbre interval as the vector between two points in a timbre space. They tested the timbre-vector hypothesis by asking listeners to compare two timbre intervals (A-B versus C-D): A, B, and C were fixed and there were various Ds presented. Listeners ranked the Ds according to how well they fulfilled the analogy: timbre A is to timbre B as timbre C is to timbre D (notated A:B :: C:D; see CD1 vector in Fig. 8.2). The ideal CD vector would be a simple translation of the AB vector in the space with A, B, C, and D forming a parallelogram (shown with dashed lines in Fig. 8.2). Ehresman and Wessel found that the closer timbre D was to the ideal point defined by the parallelogram model, the higher the ranking.

McAdams and Cunibile (1992) subsequently tested the vector model using the three-dimensional space from Krumhansl (1989) and varying the orientation and length of the vector compared to the ideal values. In Krumhansl's timbre-space model, each sound had a position in the three shared dimensions, but they also had a factor specific to each sound that increased its distance from the other sounds, called its "specificity" (see McAdams, Chap. 2). The specificities were ignored in McAdams and Cunibile's calculations. They selected different kinds of Ds (see Fig. 8.2): D1 was near the ideal spot; D2 was about the same distance from C, but was at least 90° in the wrong direction; D3 was in about the right direction from

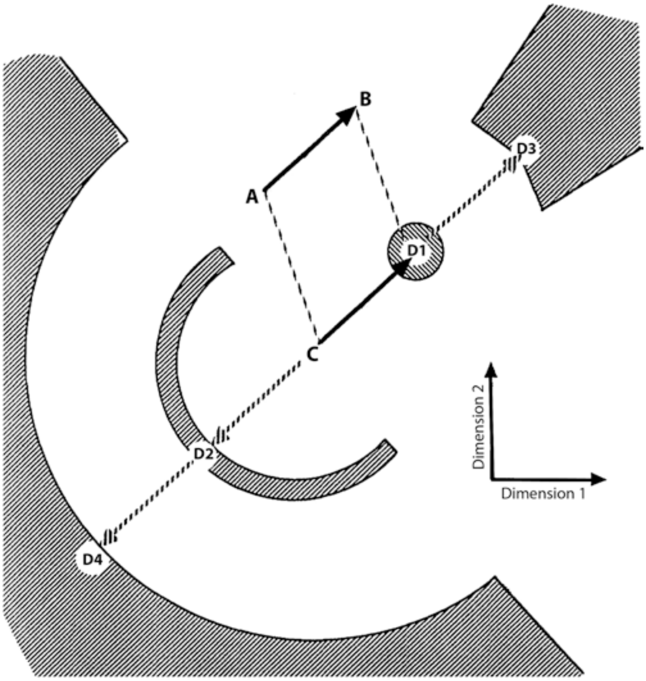

Fig. 8.2 Two-dimensional representation of the different sequence types used by McAdams and Cunibile (1992). The *hashed areas* represent the constraint space for the end points of CD vectors, which are labeled *D1*, *D2*, *D3* or *D4*, accordingly. The ideal point would be at the tip of the arrowhead for the CD vector that forms a parallelogram with AB (connected by dashed lines). For the three-dimensional case, the area would be a small sphere for D1, a shell for D2, part of a cone for D3, and a solid with a hemispherical hollow for D4. (Adapted from figure 2 in McAdams and Cunibile 1992; used with permission of The Royal Society)

C, but its length was at least 1.8 times greater than that of the ideal vector; and D4 was both too far and in the wrong direction. Listeners compared pairs of A:B :: C:D analogies constructed in this way and had to choose for which one C:D seemed most similar to A:B (e.g., compare A:B :: C:D1 with A:B :: C:D4). Five sets of timbres at different places in Krumhansl's timbre space were chosen for each comparison to test for the generality of the results. In general, timbres close to the ideal point predicted by the vector model were preferred as better fulfilling the A:B :: C:D analogy than were timbres that were at some distance from that point. Both nonmusicians and composers of electroacoustic music found the task rather difficult. This shouldn't be too surprising given that even professional composers have had almost no experience with music that systematically uses timbre intervals to build musical structures. Support for the model was stronger for electroacoustic composers than for nonmusicians, however, suggesting some effect of musical training and experience.

When one examines in detail the five different versions of each comparison type, it is clear that not all timbre comparisons go in the direction of the model predictions. One confounding factor is that the specificities on the sounds in this set were ignored in computing the vectors and selecting the analogies. These specificities would of course distort the vectors that were used to choose the timbres because they are like an additional dimension for each timbre. As such, certain timbre intervals corresponded well to what was predicted because the specificities were absent or low in value, whereas others were seriously distorted and thus perceived as less similar to other intervals due to moderate or high specificity values.

The relative lack of generalizability of timbre interval perception across different timbres may be due to a number of factors that were not controlled in McAdams and Cunibile's study. First, there may be a relative instability of judgement strategies given that most listeners have never encountered a listening situation in which perceiving abstract timbral relations was appropriate. Second, there may be effects of the relative magnitude of a given vector and the distance between to-be-compared vectors: it may be difficult to compare with precision very large vectors or small vectors that are very far apart in the space. What this line of reasoning suggests is that the use of timbre intervals as an integral part of a musical discourse runs the risk of being very difficult to achieve with very complex and idiosyncratic sound sources, such as acoustic or electronic musical instruments, because they will, in all probability, have specificities of some kind or another.

It may be difficult to use timbre intervals as an element of musical discourse in a general way in instrumental music given that the timbre spaces of acoustical instruments also tend to be unevenly distributed (see Fig. 8.1), unlike the regular spacing of pitches in equal temperament. If timbre intervals are to be used, in the long run they will most likely need to be limited to synthesized sounds or blended sounds created through the combination of several instruments. Whether or not specific intervals are precisely perceived and memorized, work in progress shows that perception of the direction of change along the various dimensions is fairly robust, which would allow for the perception of similar contours (patterns of relative change along the different dimensions) in trajectories through timbre space. Indeed, McDermott et al. (2008) have shown that timbral brightness contours (patterns of ups and downs) can be recognized irrespective of the exact amount of change and also can be compared to contours in pitch and loudness.

It should be noted, nonetheless, that in a context in which pitch is a structuring factor, timbre may have difficulty imposing itself as a dominant parameter in terms of relational perception, primarily due to a sort of dominance hierarchy favoring duration and pitch relations (rhythm and melody) when several parameters are in play. Research on the conditions under which a given musical parameter plays a significant role in the perceptual structuring of music when varied in the presence of other parameters is limited and rarely goes beyond the royal couple of pitch and duration. A first attempt in this direction, which only used nonmusical sequences, was conducted by Krumhansl and Iverson (1992). They found that the classification of pitches (high versus low) or timbres (bright versus dull) was symmetrically affected by uncorrelated variation in the other parameter: reaction times for pitch were slower when having to ignore random changes in timbre compared to when timbre was held constant and vice versa. This result suggests that it is difficult to ignore either parameter (pitch or timbre) when both are changing and indicates a tight relation between timbral brightness (change in the spectral centroid) and pitch height. This link would be coherent with underlying neural representations that share common attributes such as a tonotopic organization of spectral distribution (for more on pitch-timbre interactions, see McAdams, Chap. 2).

In two other experiments, Krumhansl and Iverson (1992) asked listeners to focus their attention on either the pitch or timbre of a single target event in a short sequence and to decide whether the same event in a second sequence was identical or different with regard to the parameter being tested. In addition, the other notes around the target event could vary either in terms of the attended parameter, the unattended parameter, or both. Globally, timbres were recognized better than pitches. A change in pitch context did not affect recognition of the target timbre and, similarly, a change in timbre context left pitch recognition unaffected. A change in pitch context strongly affected recognition of the pitch of the target event, however, indicating that listeners code relations between pitches (i.e., relative pitch) in memory rather than the absolute pitch value. To the contrary, the effect of variation in timbre context only weakly affected target timbre recognition and only when there was no variation in pitch context. This result suggests that although pitch is coded in relative terms, timbre is more likely to be coded absolutely as a sound source category, and relations among timbres are only coded when pitch does not vary. Krumhansl and Iverson (1992) concluded that relational structures among timbres would be difficult to perceive in the case in which other musical parameters vary independently. Siedenburg and McAdams (2018) presented converging evidence regarding the interference of concurrent pitch variation in a timbre-sequence recognition task (also see Siedenburg and Müllensiefen, Chap. 4). It remains to be seen, however, what the interplay of pitch-based and timbre-based structuring forces would be in instrumental and electroacoustic music that is based more on sound colors and textures and less on melody and harmony.

8.3 Timbre and Auditory Grouping

Figure 8.3 summarizes the grouping processes involved in auditory scene analysis, as well as the resulting perceptual attributes related to orchestral effects produced in music. *Concurrent grouping* determines how components of sounds are grouped together into musical events, a process referred to in psychology as *auditory fusion*. This grouping process precedes, and thus conditions, the extraction of the perceptual attributes of these events, such as timbre, pitch, loudness, duration, and spatial position. The result of combining sounds concurrently in orchestration is *timbral blend* when events fuse together or *timbral heterogeneity* when they remain separate (see Sect. 8.3.1). *Sequential grouping* connects these events into single or multiple auditory streams on the basis of which perception of melodic contours and rhythmic patterns is determined (McAdams and Bregman 1979). The orchestral effect is the integration or segregation of events into streams, textures, and foreground and background layers (see Sect. 8.3.2). And finally, *segmental grouping* affects how events within streams are chunked into musical units, such as motifs, phrases, themes, and sections.

Timbral similarity contributes to the unification of segments of music that are set off from adjacent segments when timbral contrast is introduced. Continuous change in timbre is used in progressive orchestration to create timbral modulations or larger-scale orchestral gestures (see Sect. 8.4). It becomes apparent here that auditory grouping processes are implicated in many aspects of orchestration practice, including the blending of instrument timbres, the segregation of melodies and layers based on timbral differences, and the segmentation of contrasting orchestral materials that results in the creation of perceptual boundaries in musical structures.

8.3.1 Timbre and Perceptual Fusion

As indicated in Fig. 8.3, timbre emerges from the perceptual fusion of acoustic components into a single auditory event, including the blending of sounds produced by separate instruments in which the illusion of a "virtual" sound source is created. The creation of new timbres through blending thus depends on the perceptual fusion of the constituent sound events. Concurrent grouping is affected by sensory cues, such as whether the acoustic components begin synchronously (onset synchrony), whether they are related by a common period (harmonicity), and whether there is coherent frequency and amplitude behavior (McAdams 1984). The coherent behavior cues are related to the *Gestalt principle of common fate*: Sounds that change in a similar manner are likely to have originated from the same source (Bregman 1990).

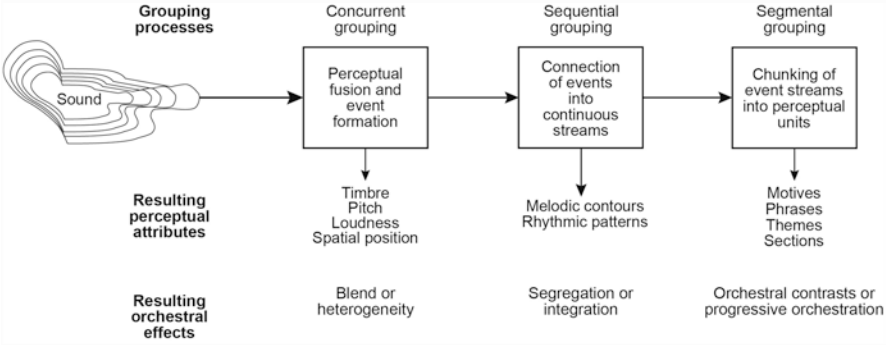
Grouping processes
Concurrent grouping
Sequential grouping
Segmental grouping
Sound
Perceptual fusion and event formation
Connection of events into continuous streams
Chunking of event streams into perceptual units
Resulting perceptual attributes
Timbre
Pitch
Loudness
Spatial position
Melodic contours
Rhythmic patterns
Motives
Phrases
Themes
Sections
Resulting orchestral effects
Blend or heterogeneity
Segregation or integration
Orchestral contrasts or progressive orchestration

Fig. 8.3 Auditory grouping processes (concurrent, sequential, segmental) that give rise to perceptual attributes and orchestral effects

Lack of synchrony, harmonicity, and parallel change in pitch and musical dynamics (*piano*, *forte*) is likely to signal the presence of two or more sound sources and to provide the information needed to organize and group their respective frequency components separately (McAdams 1984). This means that instrument combinations are more likely to be perceived as blended if they adhere to these principles. This approach can be found in orchestration manuals that suggest strict doubling of a melodic line, most often at the unison or octave intervals at which maximum coincidence of the frequency components would occur. A classic example of the combined use of these cues is Ravel's piece *Boléro*, in which he constructs virtual sound sources with complex timbres by stacking instruments in a harmonic series (fundamental pitch, octave, twelfth, double octave, etc.) and having them play in synchrony and in perfect parallelism in terms of both pitch and dynamics (https://www.youtube.com/watch?v=dZDiaRZy0Ak&frags=pl%2Cwn, instrument combinations start at 4:35 with a trumpet-flute combination or at 7:03 with French horn on the fundamental pitch and celesta and piccolos on pitches corresponding to harmonics 2–5).

The degree of fusion also depends on spectrotemporal relations among the concurrent sounds. Some instrument pairs can still be distinguished in dyads with identical pitches and synchronous onsets because their spectra do not overlap significantly. Sandell (1995) has demonstrated that sounds blend better when they have similar attack envelopes and spectral centroids, as well as a lower composite spectral centroid. He submitted listeners' blend ratings (taken as a measure of proximity) to multidimensional scaling and obtained a "blend space" whose dimensions were correlated with attack time and the spectral centroid, suggesting that the more these parameters were similar for the two combined sounds, the greater their blend. Kendall and Carterette (1993) found a similar trend concerning the role of spectrotemporal similarity in blend for wind instrument combinations. Tardieu and McAdams (2012) reported that greater blending is achieved with lower spectral centroids and slower attacks for combinations of pitched impulsive and sustained sounds (e.g., a vibraphone and bowed cello). However, the timbre resulting from the blend is determined primarily by the attack of the impulsive sound and the spectral envelope of the sustained sound, which create a chimeric sound with the head of one and the tail of the other, respectively.

In addition to global descriptors, such as a spectral centroid, research has also been conducted on the role of local descriptors of *formant structure* (prominent spectral maxima that are invariant with respect to pitch change) on wind and brass instrument blends (Reuter 2003). Lembke and McAdams (2015) extended this approach, characterizing wind-instrument spectra in terms of pitch-generalized spectral envelope descriptions. Some instruments exhibit a formant-like structure with prominent spectral peaks. They conducted two experiments employing blend-production and blend-rating tasks and studied the perceptual relevance of these formants to the blending of dyads composed of a recorded instrument sound and a parametrically varied synthesized sound. Relationships between formant center frequencies influenced blend critically, as did the degree of formant prominence.

These relations of spectral overlap and perceptual fusion seem to hold not only for the blending of sounds of pitched instruments, as mentioned above, but also for vocal sounds (Goodwin 1980) and even noise and pitched sounds as in the case of Burundian Inanga *chuchoté* (whispered Inanga) (https://www.youtube.com/watch?v=94TGtf7PdyE&frags=pl%2Cwn). The Inanga is an African zither that is traditionally accompanied by whispered voice (Fales and McAdams 1994). The latter case is fascinating because the language of that culture is tonal, and the musical instrument helps to communicate the pitch contours that disambiguate the reduced phonetic information provided by the whispered voice. These results taken together demonstrate the importance of spectral overlap in the perception of blend.

Sandell (1995) has proposed three possible perceptual results of instrument combinations. The first is *timbral heterogeneity*: individual sounds are segregated and identifiable. The second is *timbral augmentation*: subservient sounds are blended into a more dominant, identifiable sound whose timbre is then reinforced or highlighted by them. The third is *timbral emergence*: all sounds are blended and unidentifiable. An inverse relation between the degree of blend and the identifiability of the constituent sounds has been documented by Kendall and Carterette (1993). Future work is needed to develop models that can predict: (1) the degree of blend from the underlying perceptual representation, (2) the timbral qualia that emerge from blended sounds, and (3) which timbres are likely to be dominant or remain identifiable in a blend.

Additional factors that also seem to play a role in blend, but which have not been studied systematically, are event duration and spectral density. Punctuated sonorities with many instruments playing across a range of pitches for short durations, as in the opening chord of Ludwig van Beethoven's Third Symphony, the 'Eroica' (https://www.youtube.com/watch?v=nbGV-MVfgec&frags=pl%2Cwn), do not provide enough time to "hear into" the sonority and analyze out the different constituent sounds. In addition, sound masses with many closely spaced pitches are similarly difficult to analyze due to auditory limits in spectral resolution, as one finds in sound mass orchestral pieces such as György Ligeti's *Atmosphères* (https://www.youtube.com/watch?v=9XfefKJRoSA&frags=pl%2Cwn), made popular through the Stanley Kubrick film *2001: A Space Odyssey*.

8.3.2 Timbre and Sequence Perception

The connection of successive sound events into a coherent perceptual message through time is referred to as *auditory stream integration*, and the separation of events into distinct messages is called *auditory stream segregation* (Bregman and Campbell 1971). Musicians would call these streams musical lines, parts, or voices. An auditory stream is a mental representation of continuous sound activity considered by the perceptual system to be emanating from a single sound source (Bregman 1990). Sequential grouping processes organize successive events into a single stream or multiple streams based on specific cues, which are closely related to *Gestalt principles of proximity* (closeness in time) and *similarity* in auditory properties, such as pitch, timbre, loudness, and spatial position (McAdams and Bregman 1979). One of the main hypotheses behind the theory of auditory scene analysis is that the auditory system operates according to a heuristic that a sequence of events produced by a single sound source will be similar in terms of spectral content (affected by pitch register, instrument, and playing effort), intensity (affected by pitch register and playing effort), and spatial position. Continuities in these cues would thus promote the integration of the events into a stream, and discontinuities would signal the presence of other sound sources, leading to the segregation of the events into different streams within which events are similar. So sequences of sounds from different instruments can be segregated on the basis of timbre (imagine a duo of violin and piano) as can sounds from a single instrument that have very different timbral characteristics, as one might find, for example, in *Nel cor più non mi sento* (*I do not feel my heart anymore*) for solo violin by Niccolò Paganini, in which bowed and plucked sounds form separate streams in counterpoint with each other (https://www.youtube.com/watch?v=OpxwHm_a_Po&frags=pl%2Cwn starting at 1:12). It is important to note that timbre covaries with pitch, playing effort, and articulation in musical instruments and so cannot be considered independently; therefore, changing the pitch or the musical dynamic also changes the timbre.

Once timbre has been formed following concurrent grouping, it plays an important role in determining whether successive sounds are integrated into an auditory stream on the basis of similarities in spectrotemporal properties or segregated into separate streams based on timbral differences that potentially signal the presence of multiple sound sources (McAdams and Bregman 1979; Gregory 1994). This process reflects the fact that individual sources do not generally tend to change their acoustic properties suddenly and repeatedly from one event to the next (for reviews, see McAdams and Bregman 1979; Chap. 2 in Bregman 1990). As the difference between timbres gets larger, the resulting stream segregation gets stronger. Because melody and rhythm are perceptual properties of sequences that are computed within auditory streams (Fig. 8.3), timbre can strongly affect the perception of these musical patterns. A clear demonstration of this principle is depicted in Fig. 8.4.

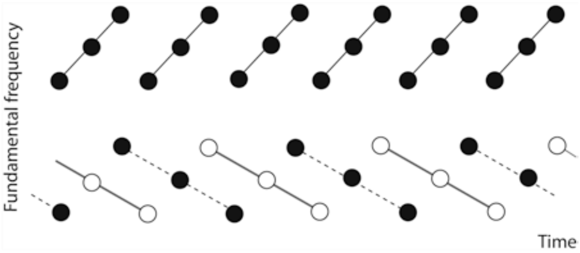

Fig. 8.4 Schematic diagram of the two versions of a melody created by David Wessel (1979) with one instrument (*top*) or two alternating instruments (*bottom*). In the *upper melody*, a single rising triplet pattern is perceived at a particular tempo. In the *lower melody*, if the timbral difference between the sounds of the two instruments (indicated by *open* and *filled circles*) is sufficient, two interleaved patterns of descending triplets at half the tempo of the original sequence are heard, as indicted by the dashed and solid lines

These early demonstrations of auditory streaming on the basis of timbre suggest a link between the timbre-space representation and the tendency for auditory streaming on the basis of the spectral differences that are created. Hartmann and Johnson (1991) argued that aspects of timbre derived from the spectral distribution are primarily responsible for auditory streaming, and temporal aspects (such as attack time) have little effect. However, several subsequent studies have indicated an important role for both spectral and temporal attributes of timbre in auditory stream segregation (Cusack and Roberts 2000; for a review, see Moore and Gockel 2002). In one study, Iverson (1995) used sequences alternating between two recorded instrument tones with the same pitch and loudness and asked listeners to rate the degree of segregation. The segregation ratings were treated as a measure of dissimilarity, and multidimensional scaling was performed to determine a *segregation space* from which acoustic properties that contributed to stream segregation could be determined. He compared the segregation space with a timbre space derived from the same sounds (Iverson and Krumhansl 1993) and showed that both static acoustic cues (such as the spectral centroid) and dynamic acoustic cues (such as attack time and spectral flux) were all implicated in segregation.

Iverson's findings were refined in an experiment by Singh and Bregman (1997). They varied the amplitude envelope and the spectral content independently and measured the relative contributions of these parameters to auditory stream segregation. A change from two to four harmonics (which would change both the centroid and the spread of the spectrum) produced a greater effect on segregation than did a change from a 5 ms attack and 95 ms decay to a 95 ms attack and 5 ms decay. Combining the two gave no greater segregation than was obtained with the spectral change, which suggests a stronger contribution of the spectral property to segregation. However, it should be noted that differences in the attacks of sounds produced by musical instruments involve many more acoustic properties than just a change in the amplitude envelope because they include noisy attack transients and rapid changes in spectral distribution during the attack.

In a slightly more musical task, Bey and McAdams (2003) used a melody discrimination paradigm. They first presented listeners with a target melody interleaved with another melody that served as a distractor such that if the two were not segregated the target melody would be camouflaged by the distractor. This mixed sequence was followed by a test melody that was either identical to the target or differed by two notes that changed the contour. Listeners were asked to decide whether the test melody was present in the previous mixture. Note that with the presentation of the test melody after the mixture, the listener must first organize the mixture into streams and then compare the melody carried by the target timbre with the ensuing test melody with the same timbre. The timbre difference between target and distractor melodies was varied within the timbre space of McAdams et al. (1995). In line with the results of Iverson (1995), melody discrimination increased monotonically with the distance between the target and the distractor timbres, which varied along the dimensions of attack time, spectral centroid, and spectral flux. Here again, the temporal and spectrotemporal properties seem to play a significant role in stream organization in addition to purely spectral properties.

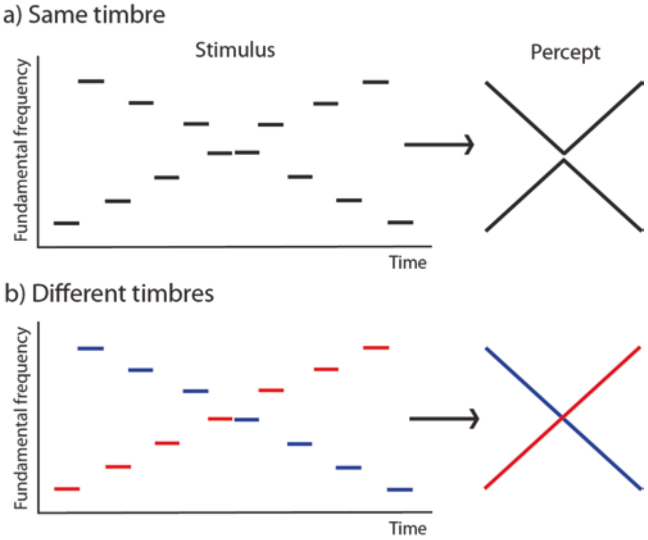

Fig. 8.5 Bouncing and crossing percepts of interleaved ascending and descending melodies depend on timbral differences. (**a**) When the timbre of ascending and descending melodies is the same, a bouncing percept is heard with V-shaped and inverted V-shaped melodies. (**b**) When the timbres are different enough (*represented by color*), the crossing melodies are heard

Timbral difference is also an important cue for following a voice (or musical part) that crosses other voices in pitch or for hearing out a given voice in a polyphonic texture with several independent parts (McAdams and Bregman 1979). Tougas and Bregman (1985) interleaved notes of ascending and descending scales, which are normally perceived as V-shaped and inverted V-shaped melodies that bounce at the crossover point (Fig. 8.5a) when the timbres of the scales are the same (same spectral content in terms of number of harmonics in their case). This percept has been interpreted as a demonstration of the role of the *Gestalt principle of pitch proximity*. However, when the spectral structures, and thus timbres, are different, this bouncing percept is replaced by the complete ascending and descending scales (Fig. 8.5b). So listeners form auditory streams with sounds of similar timbres and segregate different timbres into distinct auditory streams. Similar results are found with continuous glides of simultaneous vocal sounds composed of diphthongs: When the timbres of the vowels were the same at the moment of crossing, a bouncing perception was heard, and when they were different, crossing was perceived (McAdams and Bregman 1979; Culling and Darwin 1993). Timbre can thus play an important role in voice leading in polyphonic music. Voice leading is the connection of successive notes in a musical line, and timbre's role in this process has been largely ignored by music theorists; one exception is Huron (2016), who discusses timbral differentiation in Chap. 8 of his book.

If a composer seeks to create *Klangfarbenmelodien* (the German term for sound color melodies) that change in instrumental timbre from note to note, timbre-based streaming may prevent the listener from integrating the separate sound sources into a single melody if the changes are too drastic. The timbre sequences would not actually have the perceptual status of a melody (understood as an integrated stream within which relations among perceptual properties of successive events can be perceived) but would instead be perceptually fragmented, resulting more in a sort of *Klangfarbenzersplitterung* (sound color fragmentation)! Humans have a predisposition to identify a sound source and follow it through time on the basis of relative similarity in pitch, timbre, loudness, and spatial position (Bregman 1990; Siedenburg and Müllensiefen, Chap. 4). Cases in which such timbral compositions work successfully have used smaller changes in timbre from instrument to instrument (e.g., Anton Webern's orchestration of Bach's Ricercar from *The Muscial Offering*, https://www.youtube.com/watch?v=2cLALT09Y0M&frags=pl%2Cw*n*) or overlapping of instruments to create a line that cross-fades from one instrument to the next (e.g., Tristan Murail's *Mémoire/Érosion*, https://www.youtube.com/watch?v=dTZDCTzUcbA&frags=pl%2Cwn). However, if pointillistic fragmentation is the composer's desired aim, significant timbre change is indeed effective in inducing perceptual discontinuity.

Goodchild and McAdams (2018) propose two other groupings that are discussed in orchestration treatises but have not yet been studied empirically. They are different from stream segregation in degree and complexity more than in kind. One is *textural integration*, which occurs when two or more instruments that feature contrasting rhythmic figures and pitch materials coalesce into a single textural layer. This is perceived as being more than a single instrument but less than two or more clearly segregated melodic lines. One might consider it as occupying a middle place between integration and segregation. The emergent property is a musical surface texture. The other one is *stratification*, in which two or more different layers of orchestral material are formed perceptually and are separated into strata of greater and lesser prominence or as foreground and background layers, with one or more instruments in each layer. Integrated textures often occupy an orchestral layer in a middleground or background position, providing a textural atmosphere. An excellent example from Bedřich Smetana's *The Moldau* (measures 187–212) is the intertwining melodies of two flutes and two clarinets in a middleground texture behind the high violin melody alternating with harp arpeggios in the foregrond and horns and other strings in the background (https://www.youtube.com/watch?v=gTKsHwqaIr4&frags=pl%2Cwn starting at 5:35). The beautifully rendered middleground texture in this passage represents the shimmering of moonlight on the Moldau River. A reasonable hypothesis is that similarity of timbre, pitch register, rhythmic patterning, and articulation within layers allow for their grouping together, and differences in these parameters between layers allow for their perceptual separation.

Huron (2016) raises an interesting issue concerning orchestration practice and the effectiveness of timbral differentiation on the segregation of contrapuntal parts. He notes that contrary to what one might expect, composers often adopt more timbrally homogeneous instrument groupings in polyphonic works, such as the extensive repertoire for string quartets, brass ensembles, vocal groups, and solo keyboards. His hypothesis is that such selections of instrumentation by composers may be related to their goal of maintaining balance among the parts because heterogeneous ensembles may present perceptual difficulties due to the differences among the instruments in terms of their acoustic power and their relative salience. Huron (2016) therefore proposes that timbral differentiation is often reserved as a device used to distinguish foreground from background layers. Finally, the elusive notion of *timbral salience*, the properties that capture one's attention and lead to the distinction between foreground and background prominence, needs to be explored empirically in musical settings, although there is some research into environmental settings (Huang and Elhilali 2017).

8.4 Timbre and the Perception of Musical Form

8.4.1 Timbral Contrasts and Segmentation

Having examined how timbre derives from concurrent grouping and plays a role in sequential grouping, let us now consider how timbral discontinuities promote segmental grouping, a process by which listeners segment musical streams into units such as motifs, phrases, themes, and sections (Fig. 8.3). People organize and make sense of continuous streams of acoustic information partly by segmenting them into events, that is, meaningful units. Event segmentation is most likely an ongoing component of everyday perception that composers use. Changes in musical features are a common cue for segmentation, and listeners will indicate segment boundaries in listening experiments if strong enough changes in pitch register, dynamics, instrumentation, and duration occur. The more each feature changes and the more features that change simultaneously, the stronger is the sense of boundary (Hartmann et al. 2016). Furthermore, events are segmented simultaneously at multiple timescales and are grouped in hierarchical fashion with groups over smaller timespans being nested within groups occupying larger timespans. This nesting makes segmentation a crucial component in the formation of a hierarchical mental representation of musical form.

An "event" is some segment of time occupied by sensory information that is conceived by a listener as being bounded by a beginning and an end. For example, notes are events that are grouped into higher-level events of rhythms and melodies, which are in turn grouped into phrases and sections. The parsing of continuously incoming sensory information into events is closely related to the process of updating working memory (the part of short-term memory concerned with immediate conscious perceptual and linguistic processing) and depends on contents stored in long-term memory. Kurby and Zacks (2008) proposed that event segmentation may arise as a side effect of an adaptive mechanism that integrates information over the recent past to improve predictions about what may arrive in the near future. When perceptual features change, it becomes more difficult to predict what will follow, and errors in prediction increase momentarily. At such points, listeners need to update their memory representations of what is actually going on. Kurby and Zacks (2008) hypothesized that two processes give rise to the subjective experience that a new event has begun: (1) the detection of a momentary increase in prediction errors created by the violation of expectations generated by a model of the current event, and (2) the updating of the event model in working memory that results from the expectation violation. One important feature of this approach is the notion that events can be organized at a range of temporal grains, from fine grained (e.g., notes) to coarse grained (e.g., phrases or sections).

Event segmentation may be accomplished contextually on the basis of the internal continuity relations in the music. Generally speaking, a segment is characterized by relative internal continuity as concerns the degree and rate of change of the perceptual properties of the musical materials being heard and by a relative discontinuity at its terminal points (for an application of these principles to music analysis, see Oliver 1967). So change creates prediction errors based on presumed continuity, and the errors in turn cause segmentation. According to the *Gestalt principle of similarity*, sounds that resemble one another are grouped together and are segmented into chunks that are bounded by acoustic dissimilarities. Gradual changes over a given time period would create a sense of continuity, whereas discontinuities promote segmentation into musical units. So musical segments are formed on the basis of similarities in register, texture, and instrumentation (i.e., timbre), and changes in one or more of these musical features signal boundaries at various levels of the musical hierarchy, depending on the cumulative degree of change among them (Deliège 1989).

In their generative theory of tonal music, Lerdahl and Jackendoff (1983) proposed a series of grouping preference rules that reflect how listeners perceptually structure musical sequences. Two *Gestalt principals of temporal proximity* and *qualitative similarity* underlie the rules, most of the latter resulting from a change or discontinuity in one or more auditory attributes, including pitch register, musical dynamics, articulation (staccato, tenuto, legato, mostly related to the duration of gaps between successive notes), note duration, and timbre. Deliège (1987) experimentally tested the extent to which listeners segmented musical phrases in accordance with these grouping rules. She found that timbral discontinuities (changes in instrument timbre) were among the changes that both musician and nonmusician listeners detect most often in short phrases.

Specific evaluation of the role that timbre plays in the segmental structuring of music is limited in music-theoretical and perceptual scholarship. Goodchild and McAdams (2018) propose several types of contrasts that are often found in the orchestral repertoire: (1) antiphonal alternation of instrumental groups in call-and-response phrase structure (antiphonal is from the Greek *antiphonos*, which means "responsive, sounding in answer"); (2) timbral echoing in which a repeated musical phrase or idea appears with different orchestrations, with one seeming more distant than the other due to the change in timbre and dynamics; (3) timbral shifts in which musical materials are reiterated with varying orchestrations, passed around the orchestra, and often accompanied by the elaboration or fragmentation of musical motifs; and (4) larger-scale sectional contrasts with major changes in instrumental forces, passing from a full orchestra to a solo violin, for example. The perceptual strengths of these different contrasts depend on the timbral changes used. Furthermore, the timbral differences may play a role in which musical materials are perceived as a call versus a response in call-response patterns, or as an original pattern versus an echoed version of that pattern. Listeners also segment large-scale sections of contemporary works on the basis of marked contrasts in instrumentation and texture (Deliège 1989). Therefore, timbral discontinuities promote the creation of perceptual boundaries, whereas continuities promote the grouping of events into coherent units at various levels of the structural hierarchy. This means that timbral changes can affect both local and global levels of formal organization in music. However, timbral changes interact with changes in other musical parameters in terms of the strength of perceived boundaries. An excellent example of a timbral shift in which a melodic pattern is passed from one instrument to another can be found in Beethoven's *Egmont Overture*, with a sequence from clarinet to flute to oboe and back to clarinet with some fragmentation of the motive in the last two iterations (Fig. 8.6) (https://www.youtube.com/watch?v=2HhbZmgvaKs&frags=pl%2Cwn starting at 4:35). This pattern is repeated three more times, creating a timbral arch each time, which is set off by a two-note punctuation by most of the orchestra each time.

More research in this area would be useful to explore various timbral connections and their perception by listeners. One potential avenue for investigation is the use of timbre to create echo effects in which a repeated pattern sounds farther away on its second occurrence. Rimsky-Korsakov (1964) mentions the notion of echo phrases in which the imitation entails both a decrease in level and an effect of distance, ensuring, however, that the original and echoing instrument or instrument combination possess "some sort of affinity" (p. 110). He cites the example of a muted trumpet as being well suited to echo material in the oboes, and flutes may echo clarinets and oboes. Aside from these suggestions, one might ask: what techniques have composers used to create timbral echoes and how do they relate to perceptual principles? There are several cues to distance perception, including sound level, direct-to-reverberant energy ratio, and spectral filtering. More distant sounds are less intense, have a lower ratio of direct-to-reverberant sound energy, and have less energy in the higher frequencies due to absorption in the air and by sur-

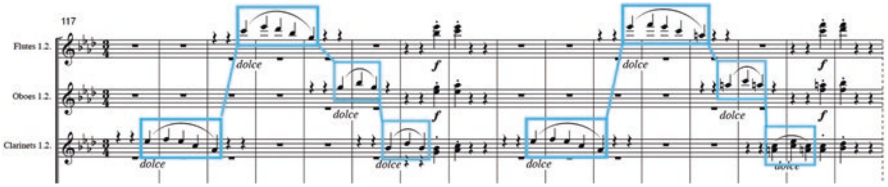

Fig. 8.6 A repeating timbral shift pattern shown by the *blue boxes* from clarinet to flute to oboe back to clarinet in Beethoven's *Egmont Overture*, measures 117–132. All instruments are notated at sounding pitch. The two-note staccato chord at the end of each pattern is also played by other instruments in the orchestra that are not shown in this detail of the score

faces in rooms (Zahorik et al. 2005). The cues that could be simulated with orchestration would be dynamics and spectral properties related to timbre (although some composers use off-stage instruments as well to get the full effect of distance).

Another process by which sequences are segmented into smaller-scale units involves detecting repeating timbral patterns and learning the transition probabilities between timbres over sufficient periods of time. Bigand et al. (1998) presented listeners with artificial grammars of musical sounds for which rules of succession of the sounds were created. After being exposed to sequences constructed with the grammar, listeners heard new sequences and had to decide if the sequence conformed to the learned grammar without having to say why. Indeed, by implicit learning of structures (language and music), any listener can know if a sequence corresponds to the structure in question without knowing why: quite simply, it doesn't "sound" right. The average correct response rate of Bigand and colleague's listeners was above chance, indicating the listeners' ability to learn a timbral grammar.

Tillmann and McAdams (2004) explored this idea further in the direction of segmentation per se based on work by Saffran et al. (1996), who sought to understand how implicit statistical learning of transition probabilities between syllables in language might lead to segmentation into words by infants. The idea is that syllables within words follow each other more often than do syllables in different words, and building up a statistical distribution of such transitions would help segment speech streams into units that correspond to words in a given language. The same research group demonstrated a similar ability in infants with pitched tone sequences, suggesting the ability applies more generally than just to speech (Saffran et al. 1999).

Tillman and McAdams applied this principal to timbre sequences using the sounds from McAdams et al. (1995) with constant pitch, loudness, and roughly equivalent duration. A lexicon of grammatical timbre triplets was created and presented sequentially in random order in an isochronous sequence for about 33 min. The probability of transition from the last sound of one triplet to the first sound of the next triplet was designed to be much lower than was the transition probability between the first and second sounds and the second and third sounds. The listeners were then tested on their recognition of triplets from the timbral grammar with the expectation that they would implicitly learn the transition probabilities among timbres. A control group was tested on a similar task without exposure to the timbral grammar. To examine the role of auditory segmentation on the basis of timbre discontinuity in the learning of timbre sequence regularities, the timbral distance relations among sounds were organized in three different conditions in which the distances between successive timbres of the triplets—as determined from the McAdams et al. (1995) timbre space—were coherent, incoherent, or neutral with respect to the grammatical triplets.

There were significant differences among the sequence types: the coherent type (S1) had the highest choice of grammatical triplets, followed by the neutral type (S3), and then the incoherent type (S2) in both learning and control groups (Fig. 8.7). So the acoustic similarities strongly affected the choices made by listeners: they preferred triplets with smaller distances between them. However, there was no

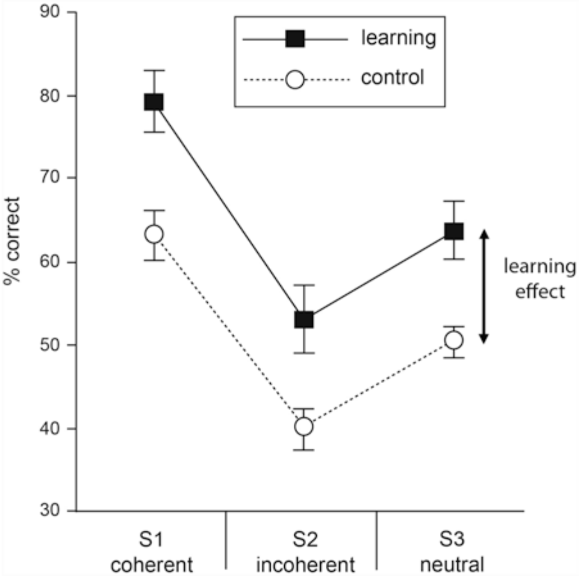

Fig. 8.7 Percentage of correct identification of timbre triplets as belonging to the timbral grammar as a function of listener group (implicit learning through hearing grammatical timbre triplets for 33 min versus control) and sequence type (*S*). The sequence types concern the alignment of acoustic grouping cues on the basis of timbral discontinuity and the implicitly learned grouping on the basis of transition probabilities (*S1: coherent*; *S2: incoherent*; *S3: neutral*). Coherent sequences were better recognized than neutral sequences, which were better than incoherent sequences. However, the amount of implicit learning was the same for all three groups. (Adapted from figure 1 in Tillmann and McAdams 2004; used with permission of The American Psychological Association, Inc.)

interaction between sequence type and listener group (the curves in Fig. 8.7 are parallel). An increase of about 14% in correct choice rate occurred for the learning group compared to the control group in all three sequence types, suggesting that learning of transition probabilities is not affected by segmentation on the basis of acoustic similarity. To summarize, even between dissimilar sounds and despite conflicting perceptual groupings, the cognitive system seems to become sensitive to statistical associative relationships among timbres. In everyday life, this capacity might be rather useful given that associations and statistical regularities (also concerning the temporal ordering of sounds) have to be learned between complex environmental sounds that can differ acoustically.

8.4.2 Timbre and Large-Scale Musical Form

Larger-scale units in music, such as formal functions (e.g., exposition, recapitulation) and types (e.g., sonata, rondo, theme, and variations), have been theorized in Classical music. Although rare, there has been some discussion of how these units can be articulated through orchestration in the music theory and musicology literature, but there is no perceptual research as yet. In classical sonata form, musical material is presented in an exposition section, then elaborated in a development section, and is returned to in a recapitulation section. Cannon (2016) studied contrasts in dynamics and orchestration (primarily instrument density) as key factors that determine whether the onset of a recapitulation serves as a resolution, climax, or arrival, on the one hand, or as a new beginning or relaunch, on the other. Examining several hundred sonata-form movements from nineteenth-century symphonies, he classified the alterations of the main theme on its return in the recapitulation into four broad types: (1) similar, (2) intensified by increased dynamic markings and instrumental forces, (3) attenuated with decreased dynamic markings and instrumental forces, and (4) contradictory with dynamics and instrumentation going in opposite directions. Cannon noted that Beethoven, for example, often intensified the theme in the recapitulation with full orchestra playing the theme at louder dynamic markings (as in the first movement of his First Symphony), a tendency observed in the majority of intensifications in Cannon's corpus. Brahms, however, often obscured the transition from development to recapitulation using lower dynamic markings. Haydn was known to use changes in orchestration in the symphonies composed during his stays in London at the end of the eighteenth century to give new color to a theme in the recapitulation (Wolf 1966). The distinction between a recapitulation being an arrival at a culminating climax point versus being a relaunch or new beginning was captured by both parameters. An arrival was often characterized by a buildup of instrumental forces and crescendo in dynamics that peaked at the onset of the recapitulation, whereas a relaunch was often characterized by a strong contrast in dynamics and instrumental texture between the end of the development and the beginning of the recapitulation. Thus, timbral factors contribute to this large-scale formal feature.

Dolan (2013) examined Haydn's structural and dramatic use of orchestration, including the process of developing variations of thematic materials. She emphasized the essential role played by orchestration in Haydn's articulation of musical form (see Chap. 2 in Dolan 2013). For example, themes that initially appear in one orchestration will return with a different one in order to nuance or transform the musical material subtly, at times keeping all other musical parameters constant, such as harmony, melody, and rhythm. Dolan describes how Haydn used opposing sonorities or textures to lend structure and dramatic impact by introducing interruptions of sections in strings or winds with full orchestral tuttis (all instruments playing together). A particularly instructive example is the second (Allegretto) movement of his "Military" Symphony, no. 100 (https://www.youtube.com/watch?v=6RmwapsXnrg&frags=pl%2Cwn). Figure 8.8 displays the evolution of instrumental involvement over the whole movement with time progressing from left to right, as indicated by the measures in the musical score on the x axis. Instrument families are shown with different hues (green for strings, blues and purples for woodwinds, orange and red for brass, and brown and black for percussion). He initially alternates sections between strings and flute, on the one hand, and single-reed and double-reed woodwinds, on the other, both occasionally punctuated with French horn

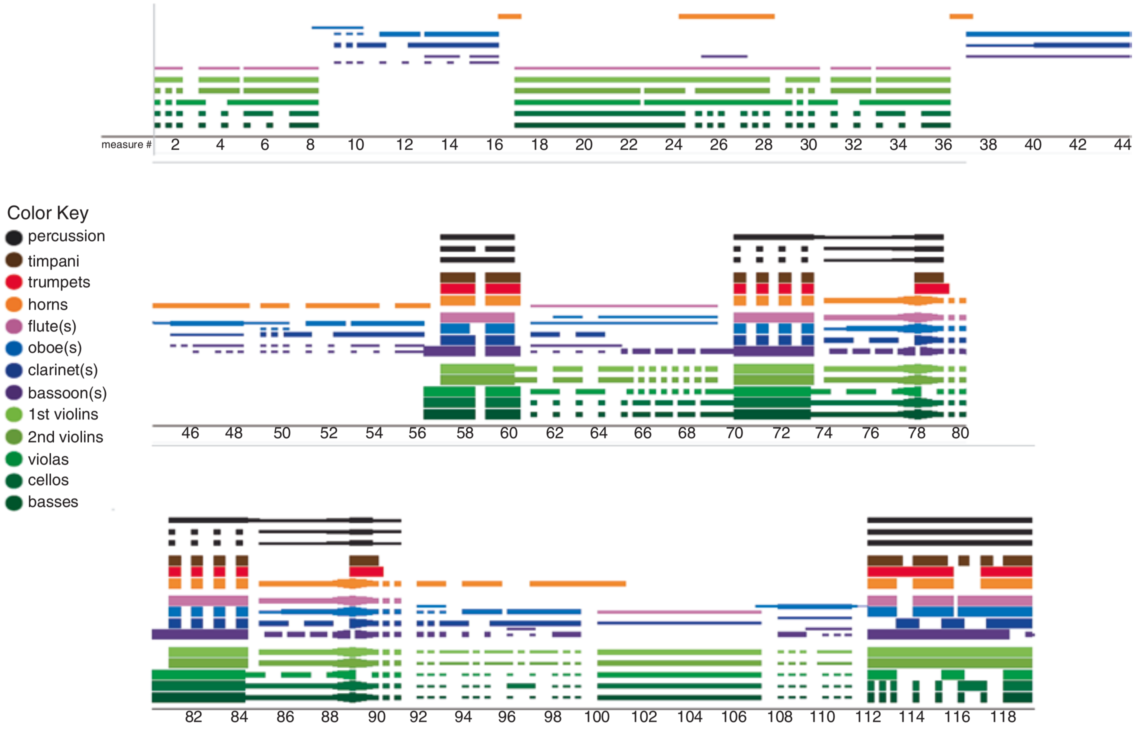

Fig. 8.8 Orchestral graph of Haydn's "Military" Symphony, no. 100, Movement II, up to measure 119. The line thickness represents the musical dynamics in the score (soft to loud). (From www.orchestralrevolution.com, figure 3.5; © 2018 by Emily Dolan, used with permission of the author)

and bassoon. Then in measure 56, a full orchestral tutti bursts in, given particular sonic power by trumpets, timpani and percussion, and the alternation between lesser and greater instrumental forces continues, which outlines the formal structure. The perceptual force of these alternations and interruptions could be tested with online segmentation techniques and perceptual characterization of the constituent sections.

Another aspect of large-scale orchestral shaping is what Goodchild et al. (2019) have called “orchestral gestures,” such as the sudden contrast between the full orchestra and a soloist or a large-scale, swelling orchestral crescendo, which contribute to peak emotional experiences in orchestral music (Guhn et al. 2007). Some orchestration treatises mention large-scale gestures on the order of a few minutes. However, a clear taxonomy of techniques and a conceptual framework related to their musical function was lacking until Goodchild (2016) developed a typology of orchestral gestures in which the time course (gradual or sudden) and direction of change (additive or reductive) in instrumentation are the primary factors. She delineated four types: gradual addition, gradual reduction, sudden addition, and sudden reduction (schematized in Fig. 8.9). These gestures are characterized by changes over time in the number and type of instruments involved, as well as in onset density by instrument family, tempo, loudness, and spectral centroid. A visualization of one of the sudden reduction excerpts from the first movement of Bruckner’s *Eighth Symphony* is shown in Fig. 8.10.

Goodchild et al.’s (2019) hypothesis was that extended patterns of textural and timbral evolution create orchestral gestures that possess a certain cohesiveness as cognitive units and have a goal-directed sense of motion. Such gestures often give rise to strong emotional experiences due to a confluence of change along many timbral dimensions, particularly timbral brightness as captured by the spectral centroid, but also changes in loudness, tempo, and registral extent (see upper panels in Fig. 8.10). Listeners’ continuous ratings of emotional intensity were recorded while listening to excerpts from the ninteenth and twentieth century orchestral repertoire.

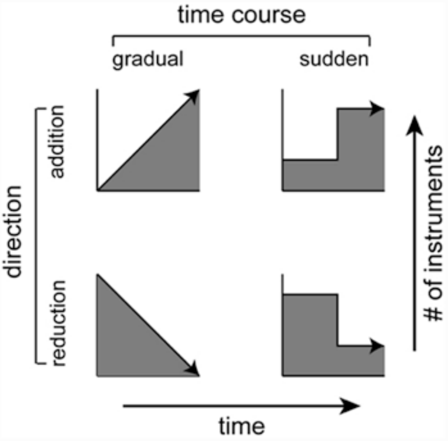

Fig. 8.9 The four types of orchestral gestures proposed by Goodchild et al. (2019). The gestures are categorized in terms of gradual or sudden change and the addition or reduction of instruments over time (From figure 1 in Goodchild et al. 2019; used with permission of Sage Publishing)

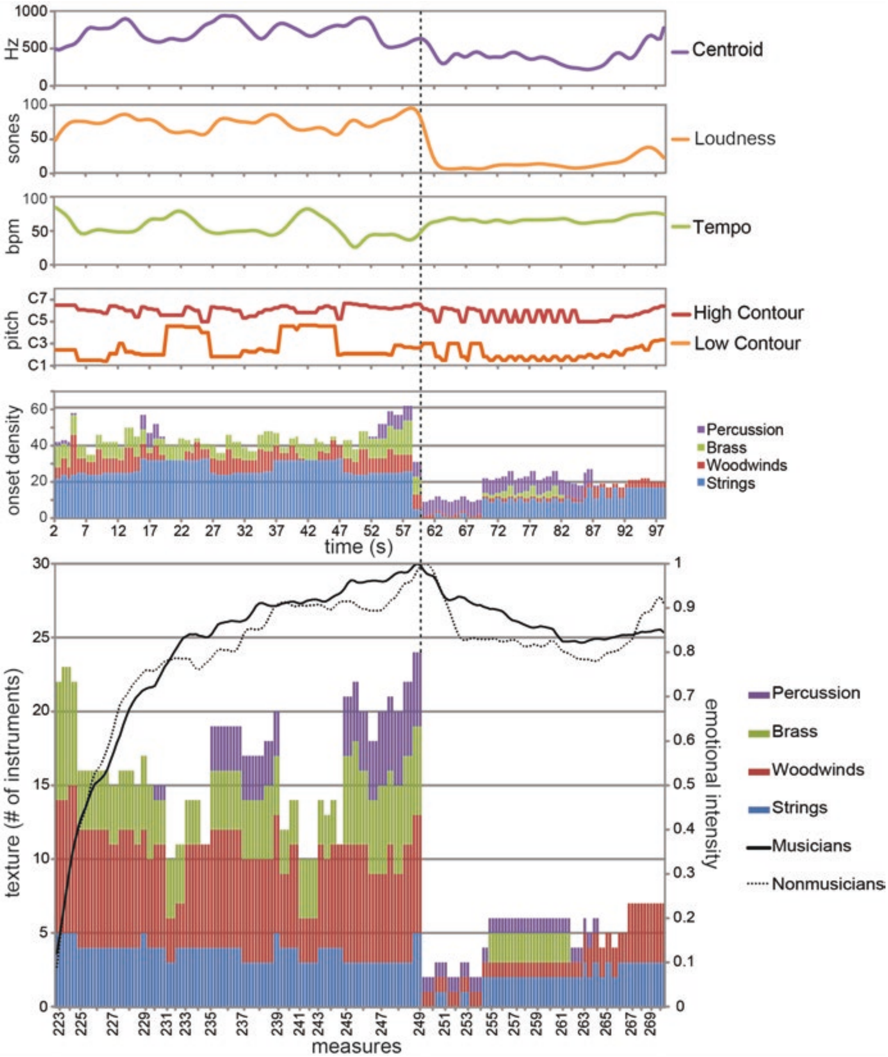

Fig. 8.10 Visualization of Bruckner's *Eighth Symphony*, first movement, measures 221–270. In the *upper panels*, spectral centroid (Hz), loudness (sones), tempo (in beats per minute), pitch range (ambitus), and onset density within instrument families are shown. The *bottom panel* graphs the instrumental texture, overlaid with the emotional intensity ratings for musician (*solid line*) and nonmusician (*dotted line*) listeners. In the bottom two panels, the *colors* represent the number of instruments of a given family that are involved. The *vertical dotted line* indicates the moment of sudden change in instrumental texture. (From figure A.11 in Goodchild 2016; used with permission of the author)

The data revealed different response profiles for each gestural type. For the gradual addition type, the emotional intensity ratings climbed steadily following the increasing growth of instrumental texture (number of instrumental parts) and loudness. For the sudden addition gestures, there was a slight tendency for musicians, but not nonmusicians, to anticipate the moment of sudden change with heightened emotional responses. Additional knowledge acquired through explicit musical training or greater experience with orchestral music may have led the musicians to develop anticipatory schemas for such changes. The responses to the gradual and sudden reductive excerpts featured a plateau of lingering high emotional intensity, despite the decrease of most musical features, including loudness, the spectral centroid (timbral brightness), instrumental texture, and onset density (number of attacks per beat). This response pattern is evident in Fig. 8.10 wherein the sudden reduction in instrumental forces at measure 250 is accompanied by a decrease in spectral centroid, loudness, and onset density along with a slight increase in tempo, and yet the average emotional intensity rating only makes a slight dip at that point.

Using re-orchestration and digital orchestral rendering as tools for testing hypotheses concerning the role of timbral brightness in emotional valence, Goodchild (2016) also showed with psychophysiological measures that the brightness of the orchestration (measured as spectral centroid) leading up to an expressive event dramatically shaped the resulting experience. Future research in this area could explore instances in which orchestral shaping (such as an abrupt change in texture or timbre) does or does not coordinate with other musical processes (such as phrase structure) to explain the interaction between formal structure based on melodic and harmonic elements and structure based on orchestration. The music-theoretical meanings and the resulting perceptual effects have yet to be explored, but this kind of work demonstrates the fertile ground that is possible in timbre cognition research through an interdisciplinary approach uniting music analysis and experimental psychology.

8.4.3 Timbre and Musical Tension

Larger-scale changes in timbre can also contribute to the expression of other higher-level structural functions in music, such as the ebb and flow of musical tension and relaxation, a type of process in music that many music theorists consider to be one of the primary bases for the perception of larger-scale form in music. When instruments composing a vertical sonority are strongly blended, timbral roughness and brightness become major components of musical tension. Nevertheless, they depend to a great degree on how the incoming acoustic information has been parsed into events and streams by auditory grouping processes. One might suppose that orchestration, in addition to pitch and rhythmic patterns, can play a major role in the structuring of musical tension and relaxation patterns that are an important component of a listener's aesthetic response to musical form. A feeling of tension accompanies a moment at which the music must continue, and a sense of relaxation signals the completion of the musical phrase or unit. In such cases, the structuring and sculpting of timbral changes and relations among complex auditory events provide myriad possibilities that composers have been exploring for decades in contemporary orchestral music but also in electroacoustic music (see Risset 2004). Musicologists have begun to address these issues, particularly as concerns timbre's potential role in what one might characterize as "musical syntax" (Roy 2003; Nattiez 2007), but psychologists have yet to tackle this area.

Experimental work on the role of harmony in the perception of musical tension and relaxation suggests that an important component of perceived tension is an attribute of timbre that is referred to as roughness (Bigand et al. 1996). The impression of timbral roughness seems to be based on the sensation of rapid fluctuations in the amplitude envelope that are correlated across peripheral auditory channels (Daniel and Weber 1997; Saitis and Weinzierl, Chap. 5). It can be generated by proximal frequency components that beat with one another. Dissonant intervals that generate an impression of roughness, like major sevenths (eleven semitones) and minor seconds (one semitone), tend to have more such beating than do consonant intervals such as octaves (twelve semitones) and fifths (seven semitones). As such, a fairly direct relation between sensory dissonance and timbral roughness has been demonstrated (cf. Plomp 1976; reviewed by Parncutt 1989).

To explore how timbre, through orchestration, might contribute to musical tension, Paraskeva and McAdams (1997) measured the effect of a change in orchestration on the inflection of tension and relaxation by comparing piano and orchestral versions of two pieces. Listeners were asked to make ratings based on the perceived degree of completion of the music at several points at which the music was stopped. What resulted was a completion profile (Fig. 8.11), which was used to infer musical tension by equating completion with release and lack of completion with tension. They tested two pieces: an excerpt from the six-voice fugue in the Ricercar from the *Musical Offering* by J. S. Bach (a tonal piece) and the first movement of the *Six Pieces for Orchestra, op. 6* by Webern (a nontonal piece, https://www.youtube.com/watch?v=NUCp4QvZxE8&frags=pl%2Cwn). Each piece was played both in an orchestral version (Webern's orchestration of the *Musical Offering* was used for the Bach piece; see link in Sect. 8.3.2) and in a direct transcription for piano of the original orchestral version of the Webern movement. Both versions were realized with a digital sampler to ensure that the performance nuances (timing, phrasing, etc.) were similar between the two. There were only very small differences between the completion profiles for musicians and nonmusicians, indicating that musical training didn't affect the completion ratings. Both tonal and atonal pieces produced significant fluctuations in musical tension, which is interesting given that some theorists feel that atonal music is devoid of this particular dimension of musical experience because it does not follow the standard tonal schemas (Lerdahl 1992). The important result here is that there were significant differences between the piano and orchestral versions, indicating an effect of timbre change on perceived musical tension. Notably, when the two versions *were* significantly different at a given stopping point (asterisks in Fig. 8.11), the orchestral version was always more relaxed than the piano version.

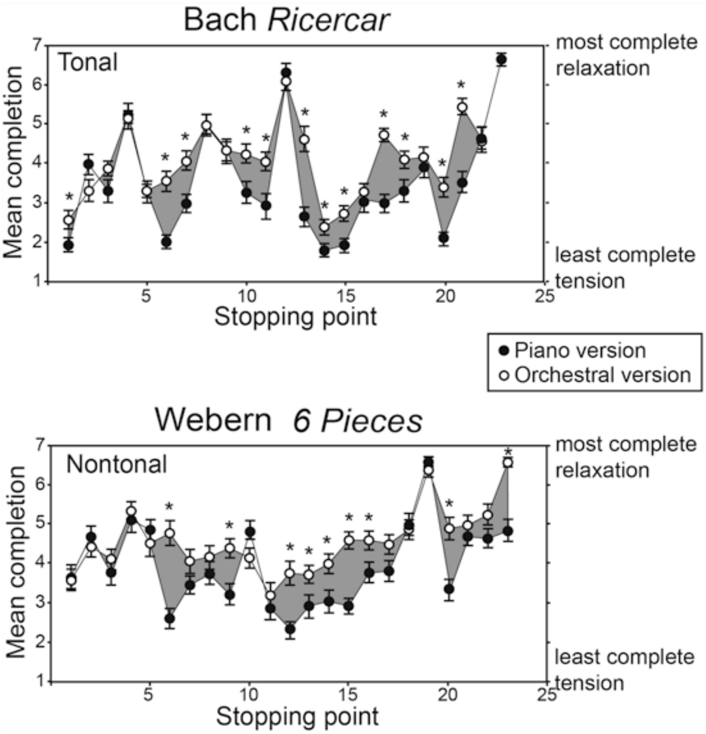

Fig. 8.11 Profiles of average completion ratings for piano and orchestral versions of tonal and nontonal pieces. The music was played to a certain point and stopped, at which point listeners made the completion rating. The next trial played the music from the beginning to the next stopping point. *Grey areas* highlight the differences between versions. *Asterisks* indicate significant differences between versions at a given stopping point. Error bars represent one standard error of the mean. (Adapted from Paraskeva and McAdams 1997; used with permission of the authors)

The hypothesis advanced by Paraskeva and McAdams (1997) for this effect was that the higher relaxation of the orchestral version might have been due to processes involved in auditory stream formation and to the dependence of perceived auditory roughness on the results of such processes. Wright and Bregman (1987), for example, illustrate several ways in which concurrent and sequential grouping processes interact and affect the perception of dissonance and tension in polyphonic music. Timbre, or any other auditory attribute of a unified sound event, is computed after auditory organization processes have grouped the bits of acoustic information together (Fig. 8.3). It may be that the same is true of sensory dissonance or auditory roughness, if we consider it to be a property of a concurrently grouped sound event. Piano sounds, being percussive in nature, have a rather sharp attack compared to most sounds from bowed and blown instruments. If several notes occur at the same time in the score and are played with a piano sound, they will be quite synchronous (particularly on a digital sampler). Because they all start at the same time and have similar amplitude envelopes and similar spectral distributions, they will have a greater tendency to be fused together. The computed roughness may then result from the interactions of all the frequency components of all the notes grouped together, although the effect of concurrent grouping on roughness perception does not seem to have been explicitly tested.

The situation is likely to be quite different for the orchestral version. For one, the same timing was used for piano and orchestra versions in the digital sampler. In the orchestral version, instruments with both slower and faster attacks were used. In other words, there is a greater range of attack times across wind and string instruments, depending on articulation, than would be the case with piano tones. Therefore, greater asynchrony could occur between the instruments in terms of perceived attack time, and the attack time difference is likely to reduce the perceptual fusion (Tardieu and McAdams 2012). Furthermore, the timbres of these instruments are often quite different. If several musical lines with different timbres arrive at the same moment on different pitches of a chord, the simultaneity may not be perceived as such because the listener may continue to track individual instruments sequentially in separate auditory streams on the basis of both the timbral similarity of notes from the same instrument and relative pitch proximity—what music theorists call voice leading (see Sect. 8.3.2).

Bregman and Pinker (1978) have demonstrated the interplay of concurrent fusion and sequential stream formation and conceived a sort of competition between the two auditory organization processes. Therefore, the attack asynchrony and the decomposition of simultaneities into separate auditory streams whose events are timbrally similar would work together to reduce the degree of perceptual fusion. A reduction in fusion would lead to greater segregation, and any roughness in the orchestral version would be computed on each individually grouped auditory event rather than on the whole harmonic complex. These individual roughnesses in the orchestral version would be much less than those of the piano version, which get grouped together more strongly. So once again, the perceptual effects of orchestration can have a very tight interaction with the processes of auditory scene analysis.

8.5 Reflections on Timbre and Musical Structure

Listeners are able to implicitly learn grammars built on rules for the probability of transition between timbres without explicit training. However, as Tillmann and McAdams (2004) demonstrated, listeners prefer certain relations among timbres that form coherent musical patterns and that distinguish among patterns. This result opens a vast field of possibilities for the construction of a veritable musical syntax based at least partially on timbre. For the syntactical use of timbre to have meaning in music, listeners must be able to learn rules of transition between timbres, as they do with durations and pitches. This learning has to be achieved implicitly by simply listening to the music without explicit training. Although the necessity of learning musical relations is obvious if one hopes to comprehend the resulting musical structures, the only explicit and experimentally controlled demonstrations of this capacity for timbre have been by Bigand et al. (1998) and Tillmann and McAdams (2004) (mentioned in Sect. 8.4.1). These findings raise the possibility of employing timbre as a primary parameter and structuring force in music.

Nattiez (2007) has critiqued Meyer's (1989) distinction between primary and secondary musical parameters and has questioned Meyer's relegation of timbre to secondary status. In Meyer's conception, primary parameters, such as pitch and duration, are able to carry syntax. (Meyer probably really meant inter-onset intervals, which define rhythm, rather than note duration, because duration per se is probably a secondary parameter related to articulation—staccato and legato.) According to this proposal, syntactic relations are based on implications for what follows next (expectations) and the realization (or not) of those implications, which is not possible with secondary parameters because they are not organized in discrete units or in clearly recognizable categories. Snyder (2000) proposes that we hear secondary parameters (including timbre) simply in terms of their relative amounts (on more of an ordinal scale), making them more useful for musical expression and nuance than for building grammatical structures.

Contrary to this position, Nattiez (2007) claims that timbre can be used to create syntactic relations that depend on expectations that lead to a perception of closure. He based his claim on his own analyses of Western and non-Western musical traditions, as well as Roy's (2003) analyses of electroacoustic music. Nattiez (2007) concluded that the main limit of Meyer's stance concerning timbre was that he confined his analyses to works composed in terms of pitch and rhythm (what current scholars of contemporary classical music call "pitch-based music"). In most cases in these styles of music, timbre is indeed only allowed to play a secondary functional role. Nattiez argued that timbre can be used to create syntactic relations that: (1) depend on expectations, leading to a perception of closure; or (2) are quite simply learned by a community of listeners as serving a given musical function within a system of hierarchical relations. He presented convincing cases supporting this hypothesis in analyses of the timbral structures in music as diverse as orchestral pieces by Debussy, Japanese drumming, and the throat singing tradition of Inuit women in northern Québec.

This debate recalls the distinction by composer and re-orchestrator John Rea between *prima facie* and *normative* orchestration (personal communication, October 26, 2011). Normative orchestration refers to situations in which the musical materials and structure are conceived in terms of pitch, harmony, duration, rhythm, and the formal structures based on them. Orchestration consists of highlighting, reinforcing, or cosmetically coloring these structures, although many orchestration decisions may be related to various programmatic topics (rapid string tremolos for storms, horn calls for the hunt or forest scenes, triumphant brass and percussion for military references) or emotional states (deep, dark melancholy of low cellos, bassoons and bass clarinets versus joyous country dance celebration with higher register woodwinds). Prima facie orchestration, to the contrary, concerns composition in which aspects of timbre are conceived at the outset as an integral part of the musical materials and forms. Examples from the electroacoustic music of Robert Normandeau, such as the piece *Tangram* (https://www.youtube.com/watch?v=KVB

RdbjQJbM&frags=pl%2Cwn), orchestral music such as *Polymorphia* by Krzysztof Penderecki (https://www.youtube.com/watch?v=9mYFKJBgxbM&frags=pl%2Cwn), or music that mixes acoustical instruments and computer-generated sounds such as *Archipelago* by Roger Reynolds are excellent examples to understand these possibilities. But even in the orchestral music of Haydn, Mozart, and Beethoven in the high Classical period, timbre plays a structuring role at the level of sectional segmentation induced by changes in instrumentation. These segmentations distinguish individual voices or orchestral layers that are composed of similar timbres and structure orchestral variations in symphonic forms (Dolan 2013).

In addition to contributing to the organization of auditory streams and orchestral layers, to contrasting materials that evoke segmentation at various levels of musical structure and to form large-scale orchestral gestures, timbre can also play a role in the ebb and flow of musical tension and relaxation and can thus contribute to the inherent expression of musical form as experienced by listeners. When instruments fuse into a musical sonority, the resulting auditory roughness, as an aspect of timbre, constitutes a major component of musical tension. However, perceived roughness strongly depends on the way the auditory grouping processes have parsed the acoustic information into events and streams (Wright and Bregman 1987) and also depends on the musical texture (homophony, polyphony, or heterophony) (Huron 2016). As a factor structuring tension and relaxation, timbre has been used effectively by electroacoustic composers such as Francis Dhomont. Roy's (2003) analyses of his music demonstrated that he employs timbre to build expectancies and deceptions in a musical context that is not "contaminated" by strong pitch structures. Roy's work implies that in a context in which pitch is a structuring factor, timbre may have trouble imposing itself as a dominant parameter as mentioned above. The interaction of musical parameters in the sculpting of the experience of musical form could be a vast and rich field if both perceptual experimentation and music analysis work together in an interdisciplinary setting to get at the essence of how orchestration—in the broadest sense of the choice, combination, and juxtaposition of sounds—actually works in the music of many styles and cultures.

A last point to consider for future experimental and musicological research on timbre concerns the crucial roles of performers and sound engineers in the final result of a sought-after timbre effect. Many parameters that affect both the timbre produced directly by an instrument (temporal and spectral properties of sound events) and the fusion of the sound of an instrument with those of other instruments (onset synchrony, pitch tuning, adjustment of timbre, and relative levels of instruments) are under the control of performers. In the end, all of these factors condition the timbre that emerges and how timbres connect sequentially and create segmental contrasts. Lembke et al. (2017), for example, showed that performers' success in achieving blend depends on both the possibilities of timbral modulation of the instrument itself (bassoon and French horn in their case, with the horn providing more room for timbral modulation) and what the role of each instrumentalist is in the musical scenario (leader or follower). Generally, followers who are trying to blend into the sound of a leader tend to darken the timbre of their instrument.

Beyond these effects, one might consider the role of sound recording and mixing, which intervene before the final result on analog or digital media. Along these lines, the composer John Rea described an experience he had in 1995 with a project of re-orchestrating Alban Berg's opera *Wozzeck* for an ensemble of twenty-one musicians in the place of the full orchestra specified by Berg (personal communication, October 26, 2011). He listened to five commercial recordings of *Wozzeck* precisely because, on the one hand, he wanted to hear how sounds fused and to see if the score presented this information in a particular way; on the other hand, he had to choose an ensemble of instruments to orchestrate the harmonies in order to best "re-present" the original score. He arrived at the following devastating conclusion: "The commercial recordings contribute to the dissimulation (the 'lie' if you will) that Art requires in order to carry on a discourse." There was indeed fusion, but it differed in each case, in each recording. It was clear that each conductor, but also each sound engineer or producer, had decided what was appropriate as a blend, as fusion, and as the projection of these qualities. Some performances of the passages in question were often in paradoxical contradiction with other performances/recordings of the same passages. To make interpretive Art always implies a confluence of several (at times conflictual?) sources of imagination and comprehension of the artistic goal.

Psychoacoustics and cognitive psychology can potentially reveal a large number of possibilities for the use of timbre in music. Composers may take profit from these scientific endeavors in the composition of their works. Music theorists and musicologists may explore, through analyses of orchestration in scores and recordings of notated music and in sound materials from electroacoustic works or field recordings of unnotated music, the ways in which composers and performers use timbre as a structuring force in music.

Acknowledgements The writing of this chapter was supported by grants from the Canadian Social Sciences and Humanities Research Council (890-2014-0008), the Fonds de Recherche du Québec—Société et Culture (SE-171434), the Canada Research Chairs Program (#950-223484), and the Killam Research Fellow Program of the Canada Council for the Arts.

Compliance with Ethics Requirements Stephen McAdams declares that he has no conflict of interest.

References

Adler S (2002) The study of orchestration, 3rd edn. W. W. Norton and Company, New York

Berlioz H, Strauss R (1948) Treatise on instrumentation (trans: Front T from 1904 German edn.). Edwin F. Kalmus, New York

Bey C, McAdams S (2003) Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. J Exp Psychol Hum Percept Perform 29(2):267–279. https://doi.org/10.1037/0096-1523.29.2.267

Bigand E, Parncutt R, Lerdahl F (1996) Perception of musical tension in short chord sequences: the influence of harmonic function, sensory dissonance, horizontal motion, and musical training. Percept Psychophys 58(1):124–141. https://doi.org/10.3758/BF03205482

Bigand E, Perruchet P, Boyer M (1998) Implict learning of an artificial grammar of musical timbres. Curr Psychol Cogn 17(3):577–600

Bregman AS (1990) Auditory scene analysis: the perceptual organization of sound. MIT Press, Cambridge, MA

Bregman AS, Campbell J (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. J Exp Psychol 89(2):244–249. https://doi.org/10.1037/h0031163

Bregman AS, Pinker S (1978) Auditory streaming and the building of timbre. Can J Psychol Rev Can Psychol 32(1):19–31. https://doi.org/10.1037/h0081664

Cannon SC (2016) Altered dynamics and instrumentation at the onset of recapitulation in the nineteenth-century symphony. Analitica-Rivista online di studi musicali 8(1) http://www.gatm.it/analiticaojs/index.php/analitica/article/view/141

Culling JF, Darwin CJ (1993) The role of timbre in the segregation of simultaneous voices with intersecting Fo contours. Percept Psychophys 34(3):303–309. https://doi.org/10.3758/BF03205265

Cusack R, Roberts B (2000) Effects of differences in timbre on sequential grouping. Percept Psychophys 62(5):1112–1120. https://doi.org/10.3758/BF03212092

Daniel P, Weber R (1997) Psychoacoustical roughness: implementation of an optimized model. Acta Acustica united with Acustica 83(1):113–123

Deliège I (1987) Grouping conditions in listening to music: an approach to Lerdahl & Jackendoff's grouping preference rules. Music Percept 4(4):325–360. https://doi.org/10.2307/40285378

Deliège I (1989) A perceptual approach to contemporary musical forms. Contemp Music Rev 4(1):213–230. https://doi.org/10.1080/07494468900640301

Dolan EI (2013) The orchestral revolution: Haydn and the technologies of timbre. Cambridge University Press, Cambridge, UK

Ehresman D, Wessel D (1978) Perception of timbral analogies. Rapport IRCAM, vol 13. IRCAM-Centre Pompidou, Paris

Fales C, McAdams S (1994) The fusion and layering of noise and tone: implications for timbre in African instruments. Leonardo Music J 4:69–77. https://doi.org/10.2307/1513183

Goodchild M (2016) Orchestral gestures: music-theoretical perspectives and emotional responses. Thesis, McGill University, Montreal, QC. http://digitool.library.mcgill.ca/webclient/DeliveryManager?pid=141286&custom_att_2=direct

Goodchild M, McAdams S (2018) Perceptual processes in orchestration. In: Dolan E, Rehding A (eds) The Oxford handbook of timbre. Oxford University Press, New York. https://doi.org/10.1093/oxfordhb/9780190637224.013.10

Goodchild M, Wild J, McAdams S (2019) Exploring emotional responses to orchestral gestures. Musicae Scientiae 23(1):25–49. https://doi.org/10.1177/1029864917704033

Goodwin AW (1980) An acoustical study of individual voices in choral blend. J Res Music Educ 28(2):119–128. https://doi.org/10.1177/002242948002800205

Gregory AH (1994) Timbre and auditory streaming. Music Percept 12(2):161–174. https://doi.org/10.2307/40285649

Guhn M, Hamm A, Zentner M (2007) Physiological and musico-acoustic correlates of chill response. Music Percept 24(5):473–484. https://doi.org/10.1525/MP.2007.24.5.473

Hartmann M, Lartillot O, Toiviainen P (2016) Interaction features for prediction of perceptual segmentation: effects of musicianship and experimental task. J New Music Res 46(2):1–19. https://doi.org/10.1080/09298215.2016.1230137

Hartmann WM, Johnson D (1991) Stream segregation and peripheral channeling. Music Percept 9(2):155–183. https://doi.org/10.2307/40285527

Huang N, Elhilali M (2017) Auditory salience using natural soundscapes. J Acoust Soc Am 141(3):2163–2176. https://doi.org/10.1121/1.4979055

Huron D (2016) Voice leading: the science behind a musical art. MIT Press, Cambridge, MA

Iverson P (1995) Auditory stream segregation by musical timbre: effects of static and dynamic acoustic attributes. J Exp Psychol Hum Percept Perform 21(4):751–763. https://doi.org/10.1037//0096-1523.21.4.751

Iverson P, Krumhansl CL (1993) Isolating the dynamic attributes of musical timbre. J Acoust Soc Am 94(5):2595–2603. https://doi.org/10.1121/1.407371

Kendall R, Carterette EC (1993) Identification and blend of timbres as a basis for orchestration. Contemp Music Rev 9(1–2):51–67. https://doi.org/10.1080/07494469300640341

Krumhansl CL (1989) Why is musical timbre so hard to understand? In: Nielzén S, Olsson O (eds) Structure and perception of electroacoustic sound and music. Excerpta Medica, Amsterdam, pp 43–53

Krumhansl CL, Iverson P (1992) Perceptual interactions between musical pitch and timbre. J Exp Psychol Hum Percept Perform 18(3):739–751. https://doi.org/10.1037/0096-1523.18.3.739

Kurby CA, Zacks JM (2008) Segmentation in the perception and memory of events. Trends Cogn Sci 12(2):72–79. https://doi.org/10.1016/j.tics.2007.11.004

Lembke S, McAdams S (2015) The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds. Acta Acustica united with Acustica 101(5):1039–1051. https://doi.org/10.3813/AAA.918898

Lembke S-A, Levine S, McAdams S (2017) Blending between bassoon and horn players: an analysis of timbral adjustments during musical performance. Music Percept 35(2):144–164. https://doi.org/10.1525/mp.2017.35.2.144

Lerdahl F (1992) Cognitive constraints on compositional systems. Contemp Music Rev 6(2):97–121. https://doi.org/10.1080/07494469200640161

Lerdahl F, Jackendoff RS (1983) A generative theory of tonal music. MIT Press, Cambridge, MA

McAdams S (1984) The auditory image: a metaphor for musical and psychological research on auditory organization. In: Crozier WR, Chapman AJ (eds) Cognitive processes in the perception of art. North-Holland, Amsterdam, pp 289–323. https://doi.org/10.1016/S0166-4115(08)62356-0

McAdams S (1989) Psychological constraints on form-bearing dimensions in music. Contemp Music Rev 4(1):181–198. https://doi.org/10.1080/07494468900640281

McAdams S, Bregman AS (1979) Hearing musical streams. Comput Music J 3(4):26–43

McAdams S, Cunibile J-C (1992) Perception of timbral analogies. Philos T Roy Soc B 336(1278):383–389. https://doi.org/10.1098/Rstb.1992.0072

McAdams S, Winsberg S, Donnadieu S, De Soete G, Krimphoff J (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res-Psych Fo 58(3):177–192. https://doi.org/10.1007/Bf00419633

McDermott JH, Lehr AJ, Oxenham AJ (2008) Is relative pitch specific to pitch? Psychol Sci 19(12):1263–1271. https://doi.org/10.1111/j.1467-9280.2008.02235.x

Meyer LB (1989) Style and music: theory, history, and ideology. University of Chicago Press, Chicago

Moore BCJ, Gockel H (2002) Factors influencing sequential stream segregation. Acta Acustica united with Acustica 88(3):320–332

Nattiez J-J (2007) Le timbre est-il un paramètre secondaire? [Is timbre a secondary parameter?]. Les Cahiers de la Société Québécoise de Recherche en Musique 9(1–2):13–24

Oliver H (1967) Structural functions of musical material in Webern's op. 6, no. 1. Perspectives of New Music 6(1):67–73. https://doi.org/10.2307/832407

Paraskeva S, McAdams S (1997) Influence of timbre, presence/absence of tonal hierarchy and musical training on the perception of musical tension and relaxation schemas. In: Rikakis T (ed) Proceedings of the 1997 International Computer Music Conference. Aristotle University, Thessaloniki, Greece, [CD-ROM]

Parncutt R (1989) Harmony: a psychoacoustical approach. Springer, Berlin

Plomp R (1976) Aspects of tone sensation: a psychophysical study. Academic, London

Reuter C (2003) Stream segregation and formant areas. In: Kopiez R, Lehmann R, Wolther AC, Wolf C (eds) Proceedings of the 5th triennial ESCOM conference. epOs-music, Osnabrück, pp 329–331

Rimsky-Korsakov N (1964) Principles of orchestration: with musical examples drawn from his own works (trans: Agate E from 1912 Russian ed.). Dover, New York

Risset J-C (2004) Le timbre [Timbre]. In: Nattiez J-J (ed) Musiques, une encyclopédie pour le XXIe siècle. Tome 2: Les savoirs musicaux. Actes Sud/Cité de la Musique, Arles, pp 134–161

Roy S (2003) L'analyse des musiques électroacoustiques: modèles et propositions [Analysis of electroacoustic musics: models and proposals]. L'Harmattan, Paris

Saffran JR, Johnson EK, Aslin RN, Newport EL (1999) Statistical learning of tone sequences by human infants and adults. Cognition 70(1):27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Saffran JR, Newport EL, Aslin RN (1996) Word segmentation: the role of distributional cues. J Mem Lang 35(4):606–621. https://doi.org/10.1006/jmla.1996.0032

Sandell GJ (1995) Roles for spectral centroid and other factors in determining "blended" instrument pairing in orchestration. Music Percept 13(2):209–246. https://doi.org/10.2307/40285694

Schoenberg A (1978) Theory of harmony (trans: Carter RE from 1911 original German publication). University of California Press, Berkeley

Siedenburg K, McAdams S (2018) Short-term recognition of timbre sequences: effects of musical training, pitch variability, and timbral similarity. Music Percept 36(1):24–39. https://doi.org/10.1525/MP.2018.36.1.24

Singh PG, Bregman AS (1997) The influence of different timbre attributes on the perceptual segregation on complex-tone sequences. J Acoust Soc Am 102(4):1943–1952. https://doi.org/10.1121/1.419688

Slawson W (1985) Sound color. University of California Press, Berkeley

Snyder B (2000) Music and memory: an introduction. MIT Press, Cambridge, MA

Tardieu D, McAdams S (2012) Perception of dyads of impulsive and sustained instrument sounds. Music Percept 30(2):117–128. https://doi.org/10.1525/Mp.2012.30.2.117

Thoret E, Goodchild M, McAdams S (2018) Timbre 2018: timbre is a many-splendored thing. McGill University, Montreal. https://www.mcgill.ca/timbre2018/files/timbre2018/timbre2018_proceedings.pdf

Tillmann B, McAdams S (2004) Implicit learning of musical timbre sequences: statistical regularities confronted with acoustical (dis)similarities. J Exp Psychol Learn Mem Cogn 30(5):1131–1142. https://doi.org/10.1037/0278-7393.30.5.1131

Tougas Y, Bregman AS (1985) Crossing of auditory streams. J Exp Psychol Hum Percept Perform 11(6):788–798. https://doi.org/10.3758/BF03205976

Wessel DL (1979) Timbre space as a musical control structure. Comput Music J 3(2):45–52. https://doi.org/10.2307/3680283

Wolf EK (1966) The recapitulations in Haydn's London symphonies. Music Q 52(1):71–79. https://doi.org/10.1093/mq/LII.1.71

Wright JK, Bregman AS (1987) Auditory stream segregation and the control of dissonance in polyphonic music. Contemp Music Rev 2(1):63–92. https://doi.org/10.1080/07494468708567054

Zahorik P, Brungart DS, Bronkhorst AW (2005) Auditory distance perception in humans: a summary of past and present research. Acta Acustica united with Acustica 91(3):409–420

Chapter 9
Timbre, Sound Quality, and Sound Design

Guillaume Lemaitre and Patrick Susini

Abstract Sound quality evaluation applies the results of timbre research to the assessment of the sound quality of manufactured products (domestic appliances, transportation, etc.). This chapter first provides an overview of one methodology. A number of acoustic descriptors reflecting perceived timbre dimensions are established and used to predict users' preference judgements. Whereas such a methodology has proven very effective, it also has some limitations. In fact, most studies only consider the pleasantness of the sounds and often overlook other potential roles of sounds in products and interfaces. In the second part, the chapter introduces sound design. Whereas sound quality evaluation merely proposes a diagnostic of the timbre of existing products, sound design aims to create or modify the timbre of product sounds to meet specific intentions. These intentions consider the pleasantness, but also several other aspects of product sounds: functionality, identity, and ecology. All these aspects are interdependent and often closely related to the temporal and timbral characteristics of the sound. The chapter continues with a discussion of the roles and practices of sound designers and introduces a set of tools that foster communication about timbre between the different participants of a sound design process. In particular, the focus is on the necessity for these participants to share a common timbre vocabulary, and the potential impact of education about sounds is considered. Finally, an important functional aspect of product sound is discussed: how to design the timbre of sounds to support user interactions with the product.

Keywords Auditory pleasantness · Product sound · Sonic interaction design · Sound evaluation · Sound function · Sound identity · Timbre descriptors

G. Lemaitre (✉) · P. Susini
STMS-IRCAM-CNRS-UPMC, Paris, France
e-mail: susini@ircam.fr

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_9

9.1 Introduction

Picture yourself driving your car. The engine purrs quietly. Then you need to overtake a slower car in front of you. The turn signal emits a satisfying mechanical tick and, at your command, the engine roars energetically; the rising engine sound (its timbre, level, pitch) matches the sporty performance that you expect of your car. As you enter a denser traffic area, passing by other vehicles creates an alternating whoosh that is quite unpleasant.

Similar to cars, most manufactured products make sounds. Some of these sounds are useful, some are annoying, some are funny, some are intrusive, and some are intriguing. Some may enhance the perceived quality of a product (the tick of a luxury watch), whereas some others are so inappropriate that they are deleterious to the overall impression of the product (the irritating hiss of a poorly fitted vacuum cleaner hose). It is therefore important for product makers to evaluate how users perceive the sounds made by the products, in other words, their sound quality. Furthermore, it is extremely useful for product designers to be able to connect the perceived quality of a product sound to measurable quantities.

Predicting the perceived sound quality of a product from quantities measured on the sound (such as timbre features) is the purpose of sound quality evaluation that is discussed in Sect. 9.2. The outcome of the methodology is an algorithm that takes a sound signal at the input and produces a numerical indicator of quality at the output. This methodology has proven very useful for a number of industrial products (cars and transportation in particular). It also has a number of limitations and, in particular, considers only one aspect of the perception of product sounds: whether they are pleasant or unpleasant.

Product sounds are not only pleasant or unpleasant, however. They serve many other purposes: they contribute to the brand image and the coherence of a product, elicit emotional reactions in users, and even have functional aspects in terms of information. As such, product designers not only want to diagnose the quality of a product sound, they also want to design its timbral and temporal characteristics to address different interdependent aspects, such as pleasure, identity, and functionality, as well as taking into account the environment in which it will be heard. As an example, most people in France associate the jingle played before any vocal announcement in French railway stations with the French national railway company (SNCF). The timbral features and temporal properties of the jingle have been specifically designed to attract the attention of users and to communicate the values of the company. In addition, this sound has been designed to be enjoyable in the complex sonic environment of railway stations. Therefore, Sects. 9.3.1 and 9.3.2 of this chapter discuss the process of sound design for the creation of a new sound and how the design process considers different aspects, such as functionality, pleasantness, identity, and ecology, and their relation to timbre. Then the problem of a common vocabulary to communicate about timbral characteristics during a project design process that involves different participants is discussed in Sects. 9.3.3 and 9.3.4. In particular, a lexicon based on previous studies, using semantic descriptions of timbre as described by Saitis and Weinzierl (Chap. 5), is proposed to help participants learn to perceive and to communicate about timbre features and temporal properties. Finally, Sect. 9.3.5 focuses on *sonic interaction design*: using sounds produced by users' interactions with the product to guide or facilitate that interaction.

Coming back to the initial car example, most drivers of manual cars change gear based on how the loudness, pitch, and timbre of the engine sound changes as they step on the accelerator. This is an example of a sonic interaction, resulting from the physical behavior of the product. These sounds could in theory also be designed and engineered by electronic means, as this is already the case in many modern vehicles, to warn pedestrians of otherwise silent electric cars or to promote economic driving.

9.2 Sound Quality Evaluation: A Critical Overview

This section describes a methodology that is common to many sound quality studies reported in the literature. Such studies have a practical goal: provide the developer of a product with a tool to measure how users will appraise the sounds of the product or prototype. More specifically, this section focuses on studies that seek to develop a model that can estimate the perceived quality of a product sound (e.g., the sounds of different vacuum cleaners) from the sound signals alone (i.e., without conducting listening tests). In such studies, the quality is defined by a single numerical value, corresponding to the average judgement of a set of typical users listening to and evaluating the sounds. A general overview of this methodology is first provided in Sect. 9.2.1, followed by a more detailed description of its different parts (an even more detailed account is provided by Susini et al. 2011). The method is illustrated in Sect 9.2.3 by one practical example: the sound quality of air-conditioning units. The limitations of such a methodology are discussed in Sect. 9.2.3.

9.2.1 The Classical Methodology of Sound Quality Evaluation: Objectivation

Most sound quality studies available in the literature follow the methodology represented in Fig. 9.1. Starting from a set of recordings of a given product or family of products (e.g., a car engine, a camera, a vacuum cleaner, etc.), the methodology has three main parts. One part (detailed in Sect. 9.2.1.1) involves characterizations of the timbre of the product sounds as a set of *sound descriptors*, that is, numerical quantities calculated from the sound signal (sound descriptors are sometimes called sound features or metrics). For another part of the sound quality study, the researcher collects judgements about the "quality" of each of these sounds using listening tests, as described in Sect. 9.2.1.2. Finally, Sect. 9.2.1.3 details the mathematical models

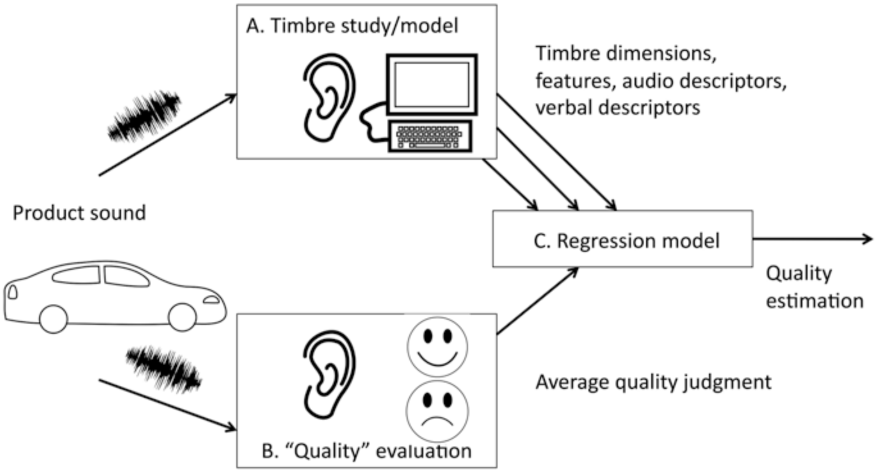

Fig. 9.1 A common methodology found in many sound quality studies consists of connecting (**A**) timbre descriptors with (**B**) quality judgements through (**C**) a regression model (Original figure)

that connect the quality judgements with the sound descriptors. Thus, the final outcome of such a procedure is an algorithm that takes the sound signal as an input and produces a quantitative indicator (a numerical value) at the output, estimating the quality of the sounds. Such an indicator is extremely valuable in an industrial context since it allows the engineers and designers to quickly evaluate how users will appraise the sounds of the product without actually collecting judgements from them with time-consuming listening tests each time a new sound is created. In this context, the whole procedure is sometimes called *objectivation*: it estimates subjective judgements (i.e., resulting from listeners' evaluations) with an objective indicator (i.e., computed from the sound signal). This methodology is rather general, and there are several variations to each part. In the next section, details are provided for each of the parts.

9.2.1.1 Characterizing the Timbre of Product Sounds

The goal of this section is to characterize each sound as a vector of numerical values that represent the listener's (i.e., the product user's) perception: the sound descriptors. Usually, only a handful of descriptors are considered. As such, sound descriptors can be considered as a low-dimensional representation of the sounds (of much lower dimensionality than the number of sound signals themselves), but they still convey the information that is important to the users in order to assess the quality of the product sounds.

Sound quality studies generally consider descriptors that are representative of a listener's perception: loudness, pitch (for sounds having pitch), duration, and timbre. As described by McAdams (Chap. 2), timbre is here considered as a multidimensional quality and characterized as a set of dimensions, each related to a sound descriptor. Product sound quality studies, therefore, use methods similar to those used to characterize the timbre of musical instruments. One such method, *semantic differentials*, consists of collecting a common vocabulary used by listeners to create scales. Another method, *dissimilarity judgements*, does not use words at all and relies instead on judgements of perceptual distances between sounds. Finally, many studies simply do not conduct any listening tests and rely on common timbre descriptors computed by software packages.

In the semantic differentials approach, the experimenters create scales, usually labeled by two opposed adjectives (e.g., clear/hazy, bright/dull): the semantic differentials (for more details, see Saitis and Weinzierl, Chap. 5). Test participants rate each sound along the set of scales and statistical techniques are used to cluster the scales into main (and often independent) factors. These main factors are interpreted and connected to the acoustic properties of the sounds, usually by listening to the sounds and picking out the best-correlated metrics.

The advantage of this method is that it allows the experimenters to generate a set of semantic descriptors of their product sounds: each main factor corresponds to a dimension of timbre that is actually perceived by the listeners and can be described by labels with a meaning shared by the listeners. In the context of sound quality evaluation, this method was used by Jeon et al. (2007), who studied refrigerator sounds with semantic differentials and identified four main factors: “booming” (and clustering pairs of adjectives such as booming/dry, trembling/flat, irregular/regular, etc.), “metallic” (metallic/deep, sharp/dull, etc.), and “discomforting” (unpleasant/pleasant, discomfort/comfort, etc.). The method, however, restricts the study precisely to what listeners can describe with words, and it is quite possible that some percepts, though perceived by listeners, may not be easily described with words. As such, the outcomes of the method strongly depend on the adjectives selected at the input in terms of relevance for describing a specific set of sounds but also in terms of the meaning of the words. This issue will be discussed in Sect. 9.3.

Instead of using a set of semantic scales, another method based on dissimilarity ratings and multidimensional scaling analysis (MDS) directly uses the ANSI definition of timbre: “the way in which musical sounds differ once they have been equated for pitch, loudness and duration” (American Standard Association 1960; Krumhansl 1989). This method was initially applied to characterize the timbre of musical instruments (Grey 1977; McAdams et al. 1995) and then of product sounds (Susini et al. 1999). In the first step of the method, listeners scale the dissimilarity between each pair of sounds from the set of product sounds under study (see Fig. 9.2).

In a second step, an MDS algorithm creates a geometrical space in which the geometrical distance between two sounds represents the perceived dissimilarity between them. The dimensions of the space are interpreted as continuous dimensions of the timbre shared by the sounds under study. As for the semantic differential method, correlations between the dimensions and the sounds’ features allows for the selection of acoustic descriptors that characterize each semantic dimension (for more detail on these two approaches, see McAdams, Chap. 2; Saitis and Weinzierl, Chap. 5).

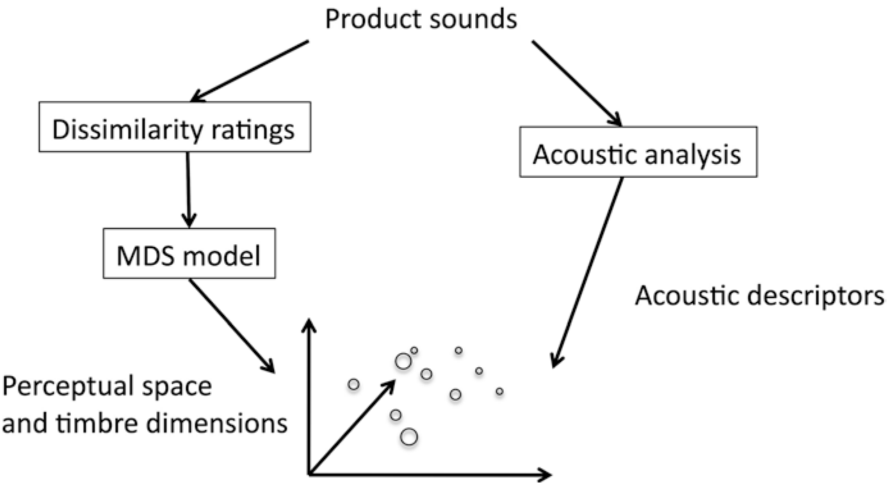

Fig. 9.2 Dissimilarity ratings and multidimensional scaling (MDS) analysis. Listeners rate the dissimilarity between the two sounds of each possible pair from a set of product sounds. An MDS model is then fit to the data and yields a perceptual space in which the geometrical distance between two sounds corresponds to the perceived dissimilarity between them. The dimensions of the perceptual space are then interpreted by correlating them with acoustic descriptors. The diameter of the circles corresponds to the position along the depth dimension, with larger circles closer to the viewer (Original figure)

The multidimensional scaling framework has the great advantage that it does not impose any predefined rating criteria on the listener. The task is thus simple and completely exploratory: as no dimension is postulated a priori, the outcome of the method (the dimensions and their descriptors) may be completely unexpected by the experimenters. This method was used by Lemaitre et al. (2007), who studied the timbre of car horn sounds and highlighted roughness, sharpness, and a third dimension specific to this set of sounds (spectral deviation), which could not have been specified without this exploratory method.

An alternative to these methods is to not conduct any listening test and simply rely instead on software packages that implement acoustic and psychoacoustic descriptors that have been found in previous timbre studies (McAdams, Chap. 2; Saitis and Weinzierl, Chap. 5). Ircam and McGill University's "Timbre Toolbox" (Peeters et al. 2011) and Lartillot and Toiviainen's (2007) "MIR Toolbox" are popular sets of Matlab functions designed for Music Information Retrieval that implement many variations of these descriptors (and see Caetano, Saitis, and Siedenburg, Chap. 11). Among commercial packages, Head Acoustics' ArtemiS (https://www.head-acoustics.de, last retrieved on July 4, 2017), Genesis's LEA (http://genesis-acoustics.com, last retrieved on July 4, 2017), and Brüel and Kjær's PULSE have been widely used in industrial contexts. These software routines are incorporated into larger program suites that also do data acquisition, analysis, and reporting and are part of the basic toolkits for many industries.

These packages include at their core a few descriptors whose calculation was formalized by Zwicker and Fastl (1990): sharpness, roughness, and fluctuation strength. *Sharpness* corresponds to a percept whereby sounds can be ordered on a scale ranging from dull to sharp or bright. It is correlated with the spectral balance of energy: sounds with more energy in low frequencies are perceived as dull whereas sounds with more energy in high frequencies are perceived as bright or sharp. *Fluctuation strength* and *roughness* both correspond to the perception of amplitude modulations in the signal, each corresponding to a different range of modulation frequencies. When modulations are slow (around 4 Hz), the sounds are perceived as fluctuating (wobbling): this is the percept of fluctuation strength. Faster modulations (around 70 Hz) are perceived as rough (harsh): this is the percept of roughness. In addition to Zwicker and Fastl's descriptors, *tonalness* (also called pitch strength or pitch salience) also plays an important role. Tonalness refers to the magnitude of the sensation of pitch in a sound (from a weak to a strong sensation of pitch) (see Hansen et al. 2011). Usually, it is estimated as the ratio of manually identified tonal components over noisy components (tone-to-noise ratio, prominence ratio) (Terhardt et al. 1982), but the perceived tonalness of sounds with multiple tonal components is still under study (also see Saitis and Weinzierl, Chap. 5).

9.2.1.2 Measuring Quality

The other important part of any sound quality study consists of collecting "quality" judgements about the product sounds. The term quality is used here in a very broad sense, as it may actually correspond to several slightly different ideas: pleasantness, unpleasantness, annoyance, merit, preference, and others. In any case, quality judgements are always obtained through a listening test and are averaged over participants into a single numerical value for each sound. Eventually, the outcome of the whole sound quality study will be an algorithm that estimates these judgements with the aim of replacing listening tests. There are two main classes of methods: sound-wise scaling procedures and paired comparisons.

In the case of *sound-wise scaling*, listeners rate each sound on a scale or a set of scales. In its simpler form, subjects rate each sound on a single scale: The annoyance of washing machine noises (Jeong et al. 2015) and the amenity (i.e., pleasantness in this context) of refrigerator noises (Sato et al. 2007) are examples of such scales. Producing absolute judgements of quality can sometimes be difficult for listeners without a context or a reference. Therefore, one variation of the method uses reference sounds to anchor the judgements. For example, Lemaitre et al. (2015a) have adapted the MUSHRA procedure (MUltiple Stimuli with Hidden Reference and Anchor, International Telecom Union 2001–2003) for collecting quality judgements of unpleasantness for wind buffeting noises. This procedure allowed the listeners to compare different sounds and then rate each sound on a scale ranging from "the least unpleasant" to "the most unpleasant." For each subset of sounds, the least and the most unpleasant sounds were systematically included to anchor the listeners' judgements.

Another difficulty of the method is to choose the label of the scale. “Quality,” “amenity,” “annoyance,” “unpleasantness,” and “noisiness” have been used, but they are also sometimes difficult to interpret for the listeners. Therefore, sets of semantic differentials (and dimensionality reduction techniques) are also used at times, just as occurs for characterizing the timbres of the sounds (see Sect. 9.2.1.1). In fact, most studies use only one listening test with semantic differential scales corresponding to both timbre dimensions and quality judgements. In one such example, the participants of a study by Hoffmann et al. (2016) rated a set of road tire noises using the following: pleasant, sharp, loud, rough, stressful, activating, and the pitch.

Another class of methods uses *paired comparisons*. The advantage is that listeners find it much easier to produce a comparison than an absolute judgement. In this case, listeners hear all possible combinations of two sounds in the set and make comparison judgements, which can be binary or on a scale. For example, they may have to select which one of the two environmental sounds is the most unpleasant (Ellermeier et al. 2004) or rate two diesel engine sounds, A and B, on a scale ranging from “I prefer A a lot” to “I prefer B a lot” (Parizet et al. 2004). The judgements for each pair of sounds are then transformed into quality scores for each sound. The simplest method consists of averaging the preference judgements for one sound across all the other sounds to which it has been compared. More complex methods rely on statistical models connecting individual quality scores to the probability of choosing sound A over B, and statistical methods are used to fit the parameters of the model. In the Bradley-Terry-Luce model (BTL), the probability of preferring sound A over B is proportional to the ratio of the quality of sound A over the sum of the quality of sounds A and B (Bradley and Terry 1952; Ellermeier et al. 2004).

Other models are derived from Thurstone’s case V model, whereby the preference probability is proportional to the cumulative normal distribution of the difference of the two quality scores (Thurstone 1927; Susini et al. 2004). One limitation of both approaches is that quality judgements are averaged across listeners and thus ignore potential individual differences in preference; indeed, two listeners can perceive the same difference between two sounds because they are different in roughness, but their preferences according to roughness can be opposed. An alternative that avoids this potential pitfall is discussed in Sect. 9.2.3. Another limitation, from a practical point of view, is that the number of pairs grows rapidly with the number of sounds. Thus, this method is limited to a rather small number of sounds and requires lengthy listening tests (see Sect. 9.2.3) unless the stimulus pairs are partitioned across subjects (Elliott et al. 2013).

9.2.1.3 Connecting Timbre with Quality

The final step of the method is to connect quality judgements with sound descriptors. Such a connection is made through a model that takes the sound descriptors as input vectors and produces estimates of quality judgements at the output. The most commonly used model is the *multivariate linear model*, whereby the quality judgement of a given sound is estimated from a linear combination of a set of descriptors. The coefficients of the model (i.e., the contribution of each descriptor to the quality of the sounds) are determined by fitting the model to the data (sound descriptors and quality judgements).

One of the difficulties of this approach is to select the descriptors that enter the model. When the timbre of the product sound has been characterized with a listening test (see Sect. 9.2.1.1), the most straightforward solution is to use the results of this initial phase as inputs to the model. When the initial timbre characterization is missing, hundreds or thousands of descriptors are often available to investigators from various software packages. Since the quality judgements result from listening tests, the set usually consists of 10–100 examples (it is very difficult for listeners to provide more judgements in a test). Using all available descriptors in the model is simply not possible because it would overfit the data; therefore, experimenters have to select a subset of them. One simple method requires the investigators to listen carefully to the sounds, consider the listeners' comments, and manually pick out the descriptors that are the most likely candidates. More systematic methods (e.g., stepwise regression, Monte Carlo) test different versions of the model to select the best subset of descriptors (Lemaitre et al. 2015a).

Linear models rely on the assumption that the contribution of each descriptor to the quality of the product sound is linear and therefore monotonic. This assumption is reasonable for loudness, as users usually prefer quieter over louder product sounds. But this is not necessarily the case for every descriptor. For example, Pietila and Lim (2015) found that listeners disliked the sounds of golf clubs hitting a ball that were too low or too high in pitch. In such cases, nonlinear models are required. One solution is to use polynomial regression or create nonlinear transformations of the descriptors within a linear regression model. These methods, however, require specifying the exact form of the nonlinearity.

All of these methods usually consider quality judgements averaged across listeners. However, different listeners may have different preferences, as illustrated by the example below.

9.2.2 Example: The Sound Quality of Air-Conditioning Units

Air conditioning units in homes, offices, and vehicles are sources of noises that can sometimes be extremely tiresome and unpleasant. Aerodynamic turbulences are created by air being blown out of a duct through a vent, and sometimes a grid, resulting in wideband noises and hisses. As a consequence, many studies have sought to quantify which timbral characteristics of these sounds are unpleasant and thus focus the efforts of engineers on reducing these unpleasant characteristics.

A study on the sound quality of indoor air-conditioning units illustrates the different steps of the methodology outlined in Sect. 9.2.1 (Susini et al. 2004). In a first step, the study characterized the timbre with dissimilarity ratings and MDS. This analysis yielded three dimensions and showed that the first dimension corresponded

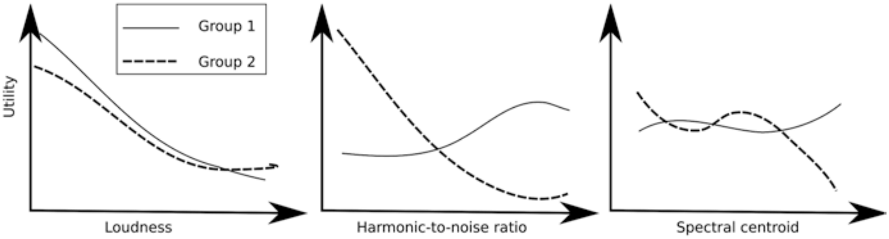

Fig. 9.3 Utility functions for the sounds of air conditioning units, three descriptors, and two latent groups of listeners (*solid line*, *dashed line*). The utility functions represent how preference judgements change with the sound descriptors. (Adapted from Susini et al. 2004; used with permission from Elsevier)

to the relative balance of the harmonic (motor) and noise (ventilator) components (harmonic-to-noise ratio). The second dimension corresponded to the spectral centroid of the sounds (a descriptor similar in spirit to sharpness; see Siedenburg, Saitis, and McAdams, Chap. 1), and the third dimension corresponded to the loudness of the sounds. The experimenters then collected paired preference judgements: listeners indicated which sound they preferred in each pair of sounds. The preference probabilities were transformed into quality values for each sound with a model based on Thurstone's case V (for details, see de Soete and Winsberg 1993). Finally, the study used a statistical model to relate the quality of the sounds to a nonlinear utility function of each sound descriptor (de Soete and Winsberg 1993).

Interestingly, the analysis found different utility functions for two latent groups of listeners (latent here means that the groups were not predetermined but resulted from the analysis itself). Figure 9.3 represents these utility functions for three descriptors (harmonic-to-noise ratio, spectral centroid, and loudness) for the two groups of listeners.

Unsurprisingly, the results showed that the quality of the air-conditioning unit is related to loudness with the same decreasing monotonic function for the two groups of listeners: listeners unanimously preferred quieter sounds. More unexpectedly, the relationship between quality and the harmonic-to-noise ratio is completely different for the two groups of listeners. Whereas listeners in the first group (solid line) preferred sounds with a strong harmonic component, listeners in the second group (dashed line) preferred sounds with a strong noisy component. The situation is somewhat similar for the spectral centroid: listeners in one group preferring sounds with a lower spectral centroid and listeners in the other group providing a rather flat response. Overall, listeners in the second group focused mainly on loudness to judge their preferences. These results clearly show that, whereas listeners perceive differences of timbre more or less equivalently, each individual may prefer different timbral characteristics. Thus, as discussed by McAdams (Chap. 2), it is very important to consider individual differences and to consider nonlinear relationships between sound quality and timbre dimensions.

9.2.3 Limitations and Issues of Classical Methodology

Over the years, the classical methodology of sound quality evaluation has been applied to a variety of industrial products and has resulted in a number of quality indicators that have been used successfully by industrial practitioners. This methodology has two main advantages. First, it characterizes the timbre of the products under consideration. As such, it provides the product makers with a good understanding of the sounds of their products and, in particular, the features of the sounds that may be deleterious to the overall quality of the products. Second, the nature of the results of such studies (a piece of software takes sound signals as an input and calculates a single numerical estimation of quality at the output) makes them very easy to integrate into a measurement chain. In fact, the results of most published sound quality studies are remarkably consistent. So one may wonder if a universal estimator of sound quality could be designed that is valid for any product sound. This impression may be the result of the limitations of the methodology. The next sections discuss these limitations.

9.2.3.1 Does Sound Quality Evaluation Still Require Experimental Characterization?

Many sound quality studies have been published over the years and some results are remarkably consistent. As a matter of fact, almost all studies have found that listeners prefer quieter over louder sounds. Another common result is that quality judgements are negatively correlated with roughness (rougher sounds are evaluated as more unpleasant than smoother ones), tone-to-noise ratio and related metrics (sounds with prominent tonal components tend to be judged as unpleasant hisses), and fluctuation strength (sounds with too much fluctuation are judged as unpleasant). Convex (i.e., u-shaped) functions that relate quality and sharpness or the spectral gravity center also have been found. Listeners tend to find sounds with prominent high frequencies (shrilling) or low frequencies (rumbling) unpleasant. Of course, specific products may deviate from general tends: typically, rough engine sounds may be appropriate for sport cars or motorcycles, and some listeners in the air-conditioning study reported in Sect. 9.2.3 preferred sounds with greater high-frequency energy.

Overall, these tendencies appear to be quite strong. The logical consequence should be that one could design a universal descriptor of sound quality, valid for any kind of sound, once and for all. In fact, Zwicker and Fastl (1990) proposed such a universal indicator: *sensory pleasantness*. They mathematically defined this indicator of sensory pleasantness as a combination of loudness, sharpness, roughness, and tonality.

There are, however, a number of issues with such an indicator. First, and as illustrated by previous examples, preference may vary from one individual to another.

Second, Zwicker and Fastl's indicator uses monotonic and separable functions to relate sensory pleasantness to different timbre descriptors, whereas previous examples have hinted that these relationships may be nonmonotonic and nonlinear in some cases. These are, however, technical difficulties that could be solved in principle. Such a universal indicator would make experimental characterization of timbre and sound quality unnecessary and thus would be a great help for product developers. But the limitations of such potential universal indicators are more conceptual and result from the concepts behind the methodology themselves. The promise of a universal indicator of sound quality based on simple timbre descriptors may in fact result from an overly restrictive perspective on sound quality and timbre. The following sections take another walk through the three parts of the methodology to discuss these concepts.

9.2.3.2 Timbre and Descriptors

Most of the discussions about the timbre of musical instruments carried out in this book (see McAdams, Chap. 2) also apply to product sound quality studies. In particular, one important limitation of the experimental methods used to characterize the timbre of product sounds is that the set of sounds under study must be homogeneous: sounds should be perceived as produced by the same source and should continuously and densely span a common acoustic space (Susini et al. 2011). When different sounds are perceived as produced by completely different sources (e.g., cats, motorcycles, and sea waves) or possess too many idiosyncratic features, listeners may become unable to provide continuous ratings of dissimilarity (in an MDS procedure), and the concept of continuous timbre dimensions is no longer relevant. In fact, sound categorization and source identification are strong cognitive processes that are not always compatible with an MDS procedure (although see McAdams, Chap. 2). To address this issue in a product sound quality study, sounds are carefully selected, and listeners are required to focus on the timbral features irrespective of the sound sources. It is also not uncommon to add synthesized sounds by homogeneously varying a few synthesis parameters or to select sounds that are closely distributed in terms of timbre characteristics.

This creates two potential issues. First, there is a risk that the sound selection may not be representative of the variability of sounds emitted by the products (ecological validity). This issue can be handled by preliminary studies seeking to explore the variability of product sounds and select samples representative of this variability (Susini et al. 2004; Parizet et al. 2008).

The second issue is circularity. Because of the established tradition of sound quality indicators and timbre descriptors, there is a tendency to select sounds that are homogeneously sampled across "classical" descriptors (e.g., sharpness, roughness, tonality, etc.). Unsurprisingly, the results of experimental characterization often yield the same timbre dimensions and descriptors used to select the sounds. In addition, the experimental characterization of timbre requires a limited number of sounds (typically about twenty sounds), and a dense and homogeneous sound selection can only span a few dimensions. This may explain why many experimental studies of product sound timbre systematically yield more or less the same three to five timbre dimensions.

One deleterious consequence is that this leads to the false belief that the timbre of any sound set can be characterized by the same three to five descriptors, conveniently implemented in software packages. However, this book beautifully illustrates a number of other phenomena contributing to the timbre of sounds. First, the timbre of a product is closely related to its identity, which cannot be easily ascribed to the combination of a few descriptors. In particular, a given product generates a large variety of different sounds created by a number of different physical sources. As discussed by McAdams (Chap. 2), the identity of a musical instrument, or for that matter of an industrial product, is precisely characterized by the variability of the sound events produced by the product or products under different modes of operation. Agus, Suied, and Pressnitzer (Chap 3) also show that the identity of a sound source cannot be completely specified by a few systematic common dimensions and descriptors. In particular, idiosyncratic features of sounds play an important role in sound recognition and, by definition, are not generalizable to other sounds.

The temporal evolution of product sounds is also very important for their identity (McAdams, Chap. 2; Caetano, Saitis, and Siedenburg, Chap. 11); yet classical descriptors usually do not take time into account. More generally, whether timbre perception can be reduced to a few common continuous and simple dimensions is an extremely important and still open question in timbre research (see the discussions in Aucouturier and Bigand 2012; Siedenburg et al. 2016), and the different chapters of this book offer different perspectives on this question. Alluri and Kadiri (Chap. 6) show that different auditory dimensions (pitch, loudness, spectral centroid, harmonicity) are specifically encoded in different cortical areas. Agus, Suied, and Pressnitzer (Chap. 3) discuss very sparse spectrotemporal representations, whereas Elhilali (Chap. 12) introduces very rich and redundant representations of sounds based on modulation encoding in the brain. Therefore, characterizing the timbre of product sounds should definitely not be considered as a problem that can be solved with standard methods dating from the early 1990s. Product sound quality as an engineering practice requires standards, but these standards should evolve with timbre research. The methods whereby sound descriptors are actually discovered blindly by deep-learning algorithms (without using the experimenters' intuitions or pre-existing knowledge) are very promising.

9.2.3.3 Connecting Quality with Timbre Descriptors

The techniques used to connect quality judgements with timbre descriptors also have a number of limitations. First, the most used method is multivariate linear regression, which, by definition, assumes a linear relationship between quality judgements and timbre descriptors. However, many examples reported above show that such an assumption is not true in general. Linear regression with nonlinear transformation of descriptors, polynomial or spline regressions, can handle nonlinearity, but the experimenters have to define the exact shape of this nonlinearity (e.g., order of the polynomials, nodes, and order of the spline functions, etc.).

Machine-learning methods can address such issues because they can learn nonlinear functions empirically, directly from the data, without the need for the experimenter to specify the nature of the nonlinearities. Pietila and Lim (2015) used a nested artificial neural network to connect preference judgements of golf club sounds to three descriptors (pitch, loudness, and timbre). The results showed complex nonlinear functions connecting preference judgements to each descriptor, which could not have been found by regression-based techniques without a priori assumptions about these functions. Another advantage of this technique is that it directly connects preference judgements to sound descriptors without the specification of an intermediate model connecting preference to quality (e.g., BTL, Thurstone's case V model).

Another issue with the connection of quality judgements to timbre descriptors is the selection of descriptors that enter the model. In most reported cases, experimenters manually select the descriptors on the basis of their own listening experience. Even when the timbre has been experimentally characterized with listening tests, there is no guarantee that listeners will use the same sound features to assess both preference and the dissimilarity between these sounds. In theory, listeners may consider that one aspect of the sounds that does not contribute much to their dissimilarity (e.g., a tiny buzz) is in fact deleterious to the quality of the sounds. Here again, machine-learning techniques can address such issues. Caetano, Saitis, and Siedenburg (Chap. 11) review deep-learning techniques that can learn sound features from the sounds themselves. For example, Dai et al. (2003) used a three-layer neural network to estimate annoyance judgements of brake squealing noises. Instead of using loudness, pitch, and timbre descriptors, the input to their neural network was a vector containing the amplitude of the spectral peaks between 2 and 18 kHz (80 values). The neural network thus estimated the annoyance judgements directly from the spectrum.

More generally, using machine-learning techniques has a huge potential for sound quality evaluation (for a review, see Pietila and Lim 2012). It should be noted, however, that the classical sound quality methodology (derived from psychoacoustics) and machine learning techniques have different philosophies as regards the generalizability of the results. On the one hand, psychoacoustics is based on inferential statistics: The quality judgements collected during a listening test are assumed to be randomly sampled for a population distribution, and the result of the procedure is the probability that the quality estimation corresponds to the average judgements in the population. On the other hand, the machine learning approach is empirical. Some empirical data (i.e., quality judgements) are used to train the model, and other empirical data are used to estimate the prediction power of the model. The upside of this approach is that it is assumption-free, and the generalizability of the model is tested on real data. The downside is that, as a consequence, it requires a large quantity of empirical data to be reliable. This is a very important issue in practice, since human listeners can only provide a few quality judgements in a reasonable amount of time. An additional downside is that the resulting network structure is difficult to interpret in behavioral and/or neuropsychological terms.

9.2.3.4 Quality Evaluation

Finally, probably the biggest issue with the classical methodology is that the notion of quality itself is extremely limited. Despite variations in the practicalities, the different methods all consist of letting listeners simply rate the quality of sounds, considered as a single notion, or indicate which of two sounds they prefer.

There are several issues with this approach. First, sounds are usually played out of context. Listeners are seated in a laboratory setting, listen to sounds over headphones or loudspeakers, and indicate their ratings on some kind of interface. This is not a particularly ecological setting. In reality, product sounds are made by a visible and tangible product, manipulated by a user aiming to do something with it, in a given situation, and in a given multisensory environment. Users have expectations about the sounds of a product, especially in relation to their previous knowledge and the other sensory aspects of the product. Playing sounds to listeners in a laboratory setting eliminates most of these aspects, which do have an important contribution to the perceptual quality of a product. As a matter of fact, many studies report that a large proportion of variance in annoyance judgements of community noise (e.g., transportation noise) is related to nonacoustic factors such as personal or socioeconomic factors (Paté et al. 2017). Therefore, it should be stressed that the methodology for sound quality studies deals only with the acoustic determinants of sound quality.

Even with this qualification, a second important aspect should be considered. The “quality” of a product sound is not something that can be ordered along a single dimension. Typically, listeners dislike louder sounds when required to judge sounds heard in a laboratory setting. But take the example of a vacuum cleaner: a completely silent vacuum cleaner would be unusable because people use the sucking noise to monitor how the vacuum cleaner is operating. Furthermore, the loudness of the vacuum cleaner is often associated with power. It is difficult for users (and buyers) to consider that a very quiet vacuum cleaner may still be powerful and deserve a high price tag. So, pleasantness, usability, and power perception may actually be examples of factors that contribute to the quality of a product and are somewhat independent.

Finally, this approach advocates for the idea that product sound quality should be addressed by considering what the sounds are used for in a given product. The legal function of car horns, for example, is to warn road users of a potential danger; accordingly, designers must create new sounds that are still recognized as car horns (Lemaitre et al. 2007). Carmakers provide electrical vehicles with exterior sounds. Such sounds must be heard in the acoustic situation of normal traffic. Accordingly, studying the quality of such sounds must consider the detectability of sounds in such an environment (Parizet et al. 2014). More generally, Sect. 9.3 discusses sound design, which considers sounds in terms of different aspects that go beyond the unique notion of quality but are still related to timbre.

9.3 From Sound Quality to Sound Design

As discussed in Sect. 9.2, most studies consider only listener's overall preference and do not consider other roles that sounds could serve. In fact, many products are strongly associated with their sounds: the sound of the Harley Davidson motorcycle is a classic example. In 2013, the Italian composer Andrea Cera designed the sound of ZOE—the electric car produced by the French carmaker Renault—to inform pedestrians of its movement on the street; its timbre is now emblematic of the car's identity and is nicely integrated into the urban sound environment. For products or devices, there are also several examples that reveal how sounds can improve useful information: monitoring in intensive care units or indicating performance for physical rehabilitation or sports activities. Thus, extending the sound quality approach, the *sound design* approach embraces the fine articulation of functionality, pleasantness, identity, and ecology of product sounds and their environment. Furthermore, sound design not only considers the diagnostic of existing product sounds; it is a process to design, engineer, or modify the dynamic and interactive timbral characteristics of product sounds to meet specific intentions or requirements defined during the development of products.

9.3.1 Sound Design: Make the World Sound Better

40 years ago, in *The Tuning of the World* (1977), R. Murray Schafer wrote about a new soundscape in which natural sounds are increasingly replaced by artificial sounds (p. 91), and "warned music educators that they would now have to be as concerned about the prevention of sounds as about their creation" (p. 98). Today sounds are widely used in a variety of products, ranging from desktop computers to mobile phone applications and from safety warnings (e.g., for hospitals, aircraft) to electric cars. These new artificial sounds are functional sounds added into our environment for specific purposes—not as decorations and not as pieces of art. The aim is to produce a sound to communicate efficient information to a user. Acoustic features such as rhythm, pitch, and loudness make an alarm audible and urgent (Stanton and Edworthy 1999). Using sounds to warn of a danger, to confirm actions, or to guide someone toward a specific direction is one kind of functional sound feedback, but sounds can also be designed to improve users' performances in terms of learning and control of a device (see Sect. 9.3.5 for a discussion on sonic interactions).

Fine-tuning acoustic features to provide efficient information using a sound, such as an alarm, could be done by an engineer based on ergonomic inputs (Stanton and Edworthy 1999) or psychoacoustic tests. In this way, the alarm sound is created based on functional recommendations. In the specific case of electric vehicles, which may be dangerous for pedestrians in urban areas because they are too quiet, the addition of sounds quickly appeared necessary to allow pedestrians to not only detect the presence of the vehicle but also its position, distance, and speed. Engineers can apply those functional recommendations in terms of intensity and fundamental frequency for the sound but, in addition, a sound designer would also shape the timbre to achieve a “harmonious solution.”

From an ecological perspective, a harmonious solution is obtained if the creation of the functional sound feedback is conceived in terms of its integration into the environment in which it will be heard as if it had always been part of it, thereby producing a feeling of pleasure or satisfaction by the intrinsic characteristics of the sound. From an industrial perspective, a harmonious solution is obtained by taking into account the brand values in order to make audible the identity of the brand (or the car) through functional sound feedback. The sound designer considers the global coherence of a new functional sound and evaluates the timbre characteristics in relation to a user’s pleasure, brand identity, and the environment of use.

9.3.2 Timbre Is a Key Element for Sound Design

Susini et al. (2014) have proposed a general definition of sound design: A sound design approach is implemented to create new sounds in order to make intentions audible in a given context of use. The sounds created are referred to as *intentional sounds*. There are different types of intentions: the first intention is to efficiently reach a goal through the sound (functionality), and the second intention is to produce a harmonious solution in the environment (ecology) that combines satisfaction of the user (pleasantness) and coherence with the product (identity). Successful sound design should be the articulation of the different intentions in order to produce new interactions through sound. Thus, a designed sound must simultaneously satisfy different aspects; the consequence is that its timbre is shaped by multiple nonindependent recommendations. Indeed, the formal aspect of a sound is already largely shaped by acoustic features that are related to functional recommendations; an alarm sound has strong spectrotemporal characteristics corresponding to its warning function, and these characteristics tend to limit the possibility to shape the sound in coherence with the environment. This is the tricky part of the process undertaken by the sound designer. Functional constraints, usually related to temporal variations in intensity or pitch, must be combined with other aspects such as pleasantness, identity, and ecology, which are more related to timbre.

Most sound quality studies seek to improve the quality of everyday existing sounds resulting from natural physical/mechanical/electrical phenomena called *nonintentional sounds*: the sound of a door lock, a car engine, an electric razor, etc. Usually, this work is done by professionals in acoustics, psychoacoustics, electronics, or mechanics. In the best-case scenario, recommendations are based on a perceptual analysis of the timbre differences and similarities between a set of sounds collected from several devices covering the full range of a product, such as the example of the air-conditioning units presented in Sect. 9.2.

But what happens when the product does not exist yet, or when the device was silent but to which it has become necessary to add sounds? Typically, an electric car is quieter than a car with an internal combustion engine. However, it is widely accepted that it is necessary to add intentional sounds to alert pedestrians and also to inform the driver about the car's state of functioning (e.g., its speed). New sounds must be imagined and created that satisfy functional constraints (e.g., detectability) as well as constraints in terms of pleasantness, identity, and ecology. The fruitful approach providing relations between users' preferences and timbre attributes based on an analysis of a collection of existing sounds (i.e., the sound quality methodology) is not useful in this case. Fortunately, practice in sound design is led by a strong creative process based on different sources of inspiration; in addition to their technical skills, sound designers are characterized by creative abilities to make sound sketches composed of different timbres, which can make all the difference in producing a successful articulation between functionality, pleasantness, and identity of a new product sound with respect to the sound environment. As has been done for science fiction movies, sound designers have to imagine and create new intentional sounds for our everyday environments.

Research on sound perception and cognition can be very informative for designing sounds. Indeed, the choice of one sound rather than another can be done arbitrarily by a sound designer; however, that does not mean that any choice would do equally well. The best choice must be an intelligent fit to human perception. Imagine a case in which one has to make a choice: is it advisable to use abstract sounds[1] (artificial tones such as beeps) rather than ecologically produced sounds (everyday sounds related to a source or an action)? Compared to abstract sounds, ecological sounds are often identifiable and thus more rapidly used in line with their function: the sound of a door lock may be used as a depiction for a "closing file" action on a computer because it will be quickly and easily understandable by users. However, it has been shown that abstract sounds work just as well after users have been exposed to them for a while and have learned to associate the sound with the function (see Agus, Suied, and Pressnitzer, Chap. 3). This finding encourages designers to propose new sounds and then to drop the old clichés (e.g., the use of a reflex camera sound for a digital camera). Such knowledge is very helpful in making a decisive choice in a sound design process.

In sound design, knowledge of timbre perception is especially important. The timbre study of air-conditioning units presented in Sect. 9.2.3 is a useful example about the relevant timbre dimensions that can be used to shape a product sound with respect to a listener's preferences. Knowledge of the relation between sound identification and timbre for different physical characteristics of everyday sounds is also fundamental in sound design in order to manipulate sound descriptors to achieve specific target sounds, for example, in terms of material (e.g., a digital sound that evokes wood or metal) or form (e.g., a sound that has the same resonance as a large plate).

Finally, knowledge of timbre dimensions with respect to the values of a brand is also very useful for the sound designer working on the sound identity of the brand.

[1] The difference between artificial and everyday sounds is defined in Sect. 9.3.5.1 (cf. earcons versus auditory icons).

For example, during co-design sessions (involving marketing professionals, sound designers, ergonomists, etc.), three of the main brand values of the French railway company (SNCF)—benevolent, simple, and efficient—were associated with several words related to timbre features (e.g., dull, round, warm, and dynamic). Those timbre features were then used as recommendations by the sound designers to propose sounds for different railway station equipment (e.g., departure boards or ticket dispensers). Verbal descriptions of timbre are fundamental in initiating the creative work of the sound designer in relation to the brand identity. This last example raises several questions related to the verbal description of timbre features in a sound design process. How do different participants involved in the sound design process communicate about timbre? Is it possible to share a common language? Are there enough specific words to describe timbre and translate them directly in terms of recommendations?

9.3.3 Communication About Sounds and Timbre: Different Strategies

Saitis and Weinzierl (Chap. 5) describe different methods or tasks to investigate verbal descriptions of sounds. In the free verbalization task, participants usually produce verbal descriptions related to their expertise and to their ability to recognize sounds. As Chion (1994) has argued, "we hear as we speak."

For everyday sounds, the most common strategy is to describe the source of the sounds ("this is the sound of a hairdryer," "it is a vacuum cleaner," "this is a trumpet") or the action that produced them ("someone is hitting a glass," "this is the sound of a string being pinched," "she is pushing a switch"). This is the *causal strategy*, which is the most intuitive way to speak about sounds for nonexperts. Descriptions are sometimes solely related to a specific meaning in a specific context or location: alarm sounds in intensive care units have a specific meaning only for the staff. This is the *contextual strategy*: Verbal descriptions are not specific to a sound's features but are more context-dependent. Finally, it seems that descriptions are seldom based on the sound itself in terms of acoustic characteristics and timbre features. This is the *reduced listening strategy*: Descriptions are directly related to the features of a sound independently of the meaning, the process that produced the sound, or its location.

For abstract sounds—artificial tones or beeps that differ in fundamental frequency, harmonic series, amplitude envelope shapes, rhythms and modulations—or sounds whose source is not easily identifiable, several strategies have been highlighted. Participants use vocal imitations (i.e., they vocally reproduce spectromorphological characteristics), onomatopoeic sounds (bleep, buzz, pop), illustrative analogies (like a fog horn, siren), and verbal descriptions of timbre (soft, high, short, tonal, dull, strident, warm). The type of strategy is mostly related to a listener's expertise in the field of sound. Indeed, sound experts will be more inclined to provide verbal descriptions related to the sound itself, such as acoustic characteristics and timbre features, than illustrative analogies. Schaeffer (1966) was perhaps the first to focus on descriptions related to a sound per se; based on a phenomenological approach, he provided several sound examples that illustrated different typo-morphological concepts such as mass, grain, and melodic profile. Although original and very detailed, Schaeffer's description remains quite complex and very few people use it today in its fullest form. However, the original idea of a lexicon illustrated with sound examples is still inspiring and challenging as a way to describe timbre (see Sect. 9.3.4).

In the sound design process, especially when it involves participants with diverse levels of expertise (from clients to sound designers), this diversity of strategies can be a serious obstacle for communication. Sound designers usually need information in terms of acoustic characteristics related to timbre features or temporal properties, but initial intentions are often expressed by the client with terms related to a context or a meaning, a function or an image (cf. contextual strategy). For example, the intention for an alarm sound in the context of a hospital could be described as "alerting but kind" rather than sound features such as long, smooth, continuous, high-pitched, loud enough, which are directly informative for the sound designers. Unfortunately, there is no common practice for talking about acoustic characteristics and timbre features, and sound designers often complain about the lack of tools to communicate about sounds in a sound design process.

Recently, in order to overcome this lack of a common language, an academic review of a large number of works dealing with verbal descriptions of timbre was performed for different kinds of sounds, from abstract to everyday sounds. Then, a lexicon of thiry-five relevant timbre descriptors (e.g., dry, bright, rough, warm, round, nasal, complex, strident) was proposed as an extension of Schaeffer's (1966) fundamental concept (Carron et al. 2017).

9.3.4 Learning to Talk About Timbre in a Sound Design Process

A standardized lexicon of specific terms to describe relevant sound features is a very promising tool for the field of sound design from a practical point of view, for example, to assist in the training of the different participants involved in a sound design project to perceive and use relevant timbre features for the design process. The lexicon also would be useful to teach pupils in a sound design or post-production course who are learning to listen to timbre features and could then describe those features with a common vocabulary. In the lexicon proposed by Carron et al. (2017), each term is presented on a computer interface, defined and illustrated by musical, vocal, environmental, abstract, and effect sounds, in order to provide a large diversity of examples. This tool has been tested and approved in different case studies in which industrial partners were involved (Carron et al. 2015). From a training perspective, a set of audio tests also has been developed to evaluate participants' understanding of the lexicon; it is a complementary and indispensable element of applying the lexicon. The tests assess whether using the lexicon may improve listeners' perception of a specific feature as well as their ability to describe sounds with only the terms of the lexicon.

One set of tests is called "from word to sound"; participants are asked to choose from among five sounds the most typical sound related to a specific term. Another set of tests is called "from sound to words"; participants have to choose three words among the list of terms provided to describe the prominent auditory dimensions of a specific sound (e.g., a continuous and warm sound with a slow attack). Results of the tests are compared with previous results of twenty well-trained participants.

During a training session, individual and collective explorations of the lexicon are alternated with the different tests. After each test, terms are discussed collectively to ensure a common understanding, and eventually there is a refinement of the sound examples provided for the lexicon. This global training ensures that participants involved in the same project have a rich and varied vocabulary that is adapted to describe a large number of timbre features and temporal properties appropriate for an important variety of sounds. This procedure is an alternative to sensory evaluation often used to reveal a list of words specific to the timbre of a set of sounds in relation to consumer preferences. The sensory evaluation requires several steps of discussion, training, and testing with a panel of experts, a process which is often very long (several weeks) and specific to a set of sounds.

9.3.5 Sounds to Support User Interaction

As discussed before, designing sounds has the potential to address the different roles of sounds in products, interfaces, places, or brands. The following sections focus on fostering user interaction. In fact, sounds are particularly well-suited to guide and facilitate interactions between users and a product. Because sounds are dynamic stimuli, they can react instantly and continuously to users' actions and gestures and thus provide users with real-time feedback of their actions. Furthermore, the tight bond between audition and motor actions makes it possible to use sounds to continuously guide and facilitate gestural interactions.

The idea of using sounds to support user interactions is relatively new in the field of sound design and has been labeled as *sonic interaction design* (Serafin et al. 2011). Sonic interaction design also introduces new research questions. One new research area is determining which timbre features best support interaction. Intuitively, ecologically produced sounds (e.g., the sound of rubbing a finger against a rough surface) seem the best candidates to support an interaction (e.g., interacting with a touch screen), because most people would know how timbre features change with changing interaction parameters (e.g., a faster rubbing gesture producing higher frequencies). In the next subsections, evidence is presented that, in fact, this first intuition may be too simplistic and that further work is needed to fully understand how gestures and timbre become associated in memory.

9.3.5.1 Sonic Interaction Design: A Rationale

Designing sonic interactions consists of using and creating sounds to design, help, or augment how users interact with a product or a machine. As such, sonic interaction design fits under the larger umbrella of *interaction design*: designing the ways users may interact with systems and computer interfaces in particular (Crampton-Smith 2007).

There is a rich history of sounds in human-computer interfaces. Computer interfaces emit a variety of different sound signals, each aiming to communicate a different message (e.g., computer starting up, different types of errors). One key question for designers, therefore, is how to convey a specific message with nonspeech sounds? One common strategy consists of using artificial tones (beeps or sequences of beeps forming a melody) and relying on an arbitrary code mapping the sounds' features (pitch, loudness, duration, timbre) to the message. The messages conveyed by artificial beeps are called "earcons" (Brewster 2009). The main disadvantage is that users must learn the mapping between sound features and meaning, in other words, the code connecting the different melodies to their meaning. William Gaver, an influential perception researcher and interaction designer, proposed another strategy that does not rely on an arbitrary mapping but is instead based on the spontaneous identification of sound sources that are called "auditory icons". The most famous auditory icon created by Gaver is probably the sound of a crumpled sheet of paper thrown into a garbage can, used as feedback to indicate file deletion, which was developed for Apple computers. In this example, the meaning of the sound results from the spontaneous identification of the sound source and a metaphor: deleting an electronic document being equivalent to physically crumpling and discarding a sheet of paper (Gaver 1986, 1989).

Similar to this example, many computer sounds rely on a desktop metaphor; however, ubiquity and mobility have drastically changed how users interact with mobile phones, tablets, and connected watches. Screen-and-keyboard interactions are simply not possible in many tiny devices, and vocal and gestural interactions thus have become more and more important and sophisticated (think of the different gestures used on mobile touch screens).

The use of sounds, however, has evolved more slowly, and most electronic devices still use simple and discreet beeps. There is, nonetheless, a great potential for sounds to guide, foster, facilitate, or augment gestural interactions. Sounds are dynamic stimuli. As such, they can very easily be made to react continuously and in real time to dynamic data, such as movement data, through some sort of model or mapping (i.e., data sonification). This idea is illustrated by Tajadura-Jiménez et al. (2014), who showed that changing the timbre of the sounds made by fingers rubbing the surface of an object changed the velocity and the pressure of the fingers, as well as the tactile perception of the surface.

In fact, audition and motor actions are tightly bound, and continuous sounds created by gestures may influence the gestures themselves. Lemaitre et al. (2009) have shown in particular that a continuous sonic feedback, reacting in real time to the users' gestures, can help them learn a fine gesture more rapidly than does visual feedback. Lemaitre et al. (2015b) also have shown that playing the sound of an action (e.g., tapping, scraping) can facilitate (when the action is congruent with the sound) or hinder (when incongruent) the subsequent execution of another action.

This idea of a continuous coupling of action and sound is exactly what happens when a person learns how to play a musical instrument. To produce a good tone, a violinist bows a string, and (particularly during training) adjusts bowing action continuously by listening to the sound that is produced. This sonic feedback guides the player's control, modifying bow speed, pressure, angle, and so forth. In fact, users also use continuous sounds when interacting with other products: most people use the change of pitch in the sound of their car engine while accelerating to decide when to manually change gears. The study conducted by Jérémy Danna and colleagues illustrates another such designed sonic interaction (Danna et al. 2013, 2015). They investigated the real-time continuous sonification of handwriting gestures to facilitate graphomotor learning. They devised a system whereby the movements of a pen on a sheet of paper changed the timbre of the resulting sound. When the pen movements were too fast, the timbre became squeaky and unpleasant. Jerky movements also resulted in unpleasant crackly sounds. Overall, the results showed that this designed sonic interaction improved the kinematics of the handwriting movements. There was, however, a lack of a long-term effect, thus raising the question of the persistence of the timbre-gesture associations in memory.

9.3.5.2 New Research Questions

The key issue of sonic interaction design is to design the timbre of sounds that can effectively support an interaction with a physical object or an interface (see also Ystad, Aramaki, and Kronland-Martinet, Chap. 13). "Supporting" an interaction may in fact correspond to several aspects. It may, for example, contribute to the aesthetics of the product, in other words, make it appealing or intriguing. Such ideas are, in fact, very close to the concepts of sound quality and sound design previously discussed. But more interestingly, supporting the interaction can also mean that the sounds produced by the users' gestures effectively help or guide the interaction. As such, evaluating the effectiveness of the sound design cannot be conducted only with the method of sound quality evaluation described in Sect. 9.2, but evaluation should mainly rely on measuring users' performances at performing a task (e.g., accuracy, speed at achieving a goal, etc.). As suggested by the above example of handwriting, another very important aspect of evaluating such designs is to assess the long-term benefits and, for example, whether gesture facilitation persists after the sonic feedback has been removed.

Because sonic interactions deal specifically with continuous sounds, static timbre descriptors, such as those previously discussed, are no longer relevant. Instead, the dynamic aspects of timbre become crucial. Coming back to the example of changing gears while driving a car, it is the change of timbre during acceleration that makes a car driver change the gear.

Sonic interaction design also poses specific challenges to the relationship between sound design and timbre dimensions. Whereas sound quality evaluation consists of mapping a user's judgements to timbre features or dimensions, designing sonic interactions consists of designing sounds that can effectively guide an action. A first intuition is that ecological sounds (i.e., sounds produced by real-world phenomena or connected to gestures via models of physical phenomena) should work best (Rath and Schleicher 2008; Lemaitre et al. 2009). However, empirical results have not confirmed this intuition so far. In particular, Lemaitre et al. (2015b) have shown that playing the sound of an action can prime that action. When participants had to respond to a vocal cue by physically tapping or scraping on a response interface (*ecological sound-gesture mapping*), playing, tapping, or scraping sounds before the cue could facilitate or hinder the gesture. But the same priming effect also occurred when the tapping or scraping gestures produced simple tones at different frequencies (*arbitrary sound-gesture mapping*). No advantage was observed for ecological mappings when compared to arbitrary mappings.

These results (as well as those of Danna et al. 2015) suggest that the pitch, loudness, and timbre of sounds and motor programs can be coupled in memory, and that this association can be so strong and bidirectional that simply playing a sound that has been produced by a gesture can influence the later execution of that gesture, no matter whether the coupling has existed for a long time (such as in an ecological association) or has just been created (such as in an arbitrary sound-gesture mapping). The question thus becomes that of how are sounds memorized together with other sensory stimuli or motor programs? Similarly, the question of whether the benefits of a sonically augmented device can persist for a long time is in fact the question of the persistence in memory of sensory-motor couplings. Memory for timbre is thus a very timely and crucial research question (see Siedenbug and Müllensiefen, Chap. 4).

9.4 Summary

Most manufactured products, objects, and electronic devices that make up our environment produce sounds at some point. Some of these sounds are useful, some are annoying, some are funny, some are intrusive, and some are intriguing. Such sounds may thus enhance the perceived quality of a product, whereas others may be so inappropriate that they are deleterious to the overall impression of a brand. It is therefore utterly important to be able to assess how product sounds are perceived and to design them with intentions. These are the purposes of sound quality evaluation and sound design, two areas of applied research and professional activity that rely heavily on research on timbre perception and cognition.

The first part of this chapter described and discussed a general methodology used to evaluate sound quality of products. This methodology has three main elements: describing the timbre of the product sounds (using experimental methods or software packages), collecting listeners' quality or preference judgements, and connecting timbre features with preference judgements through various sorts of regression techniques. This method has been used extensively in a variety of applied studies and has proved to be a valuable tool for engineers, as it eventually produces an algorithm taking a sound signal at the input and producing a single indicator of quality at the output. Such quality indicators are powerful, simple to use, and do not require costly user testing.

There are, however, a number of limitations to this set of methods. First of all, they are based on a definition of timbre that may be oversimplified: it considers only homogeneous sets of sounds, a limited number of timbre dimensions, and it does not consider the sound source identities, sound idiosyncrasies, or individual differences among listeners. Furthermore, this methodology relies on a very narrow conception of the quality of a set of sounds. Most methods only consider listeners' preferences or one-dimensional ratings of quality of the sounds without taking into account the context of use.

The second part of this chapter discussed the sound design approach. In contrast to sound quality evaluation, sound design embraces a wider range of aspects of intentional sounds: functionality, pleasantness (listeners' satisfaction), identity (coherence between sound and product), and ecology (integration with the context of use).

Because of its wider scope, sound design faces a number of methodological issues. One such issue is that sound designers have to interact with many different practitioners in a company, with varying levels of sound expertise, and communication clearly is an issue. These issues are also research challenges. Chief among them is the question of how to standardize verbal descriptions of sounds. This chapter thus discussed the results of research that has studied how people talk about sounds. Most people describe the sources of the sounds, the actions that produce the sounds, the context and the location, or they produce vocal imitations, but rarely do respondents talk about timbre! The chapter then described a tool that has been designed to teach and train stakeholders of sound design projects to understand, share, and use technical descriptions of timbre (such as dry, bright, rough, warm, round, nasal, complex, strident), but also to describe the temporal characteristics (such as constant/fluctuating, ascending/descending, discontinuous, etc.) and the general qualities (such as soft/loud, low/high, short/long, etc.).

Finally, the chapter focused on one particular aspect of sound functionality: sonic interaction. In fact, recent research has shown a tight bond between audition and motor behaviors. Sonic interaction design seeks to exploit this close connection by designing sounds to help, guide, foster, or augment how users interact with a product or a machine. This approach is very promising, especially for motor rehabilitation applications, but the field of sonic interaction faces important research challenges. In particular, the dynamic aspects of timbre, timbre memory, and the nature of the interactions of sound representations with other sensory or motor modalities are important areas that contemporary research on timbre and other chapters of this book explore.

Compliance with Ethics Guillaume Lemaitre declares that he has no conflict of interest.

Patrick Susini declares that he has no conflict of interest.

References

American Standard Association (1960) USA acoustical terminology S1.1–160. American Standard Association

Aucouturier JJ, Bigand E (2012). Mel Cepstrum & Ann Ova: The difficult dialog between MIR and music cognition. In: Proceedings of the international conference on music information retrieval (ISMIR), Porto, Portugal, 2012

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika 39(3/4):324–345

Brewster S (2009) Non-speech auditory outputs. In: Sears A, Lacko JA (eds) Human-computer interaction: fundamentals. CRC Press, Boca Raton, p 213

Carron M, Dubois F, Misdariis N, Susini P (2015) Définir une identité sonore de marque: méthodologie et outils (Defining a brand's sonic identity: methods and tools). Acoustique et Techniques 81:20

Carron M, Rotureau T, Dubois F, Misdariis N, Susini P (2017) Speaking about sounds: a tool for communication on sound features. J Design Res 15(2):85–109

Crampton-Smith G (2007) Foreword: What is interaction design. In: Moggridge B (ed) Designing interactions. The MIT Press, Cambridge, MA, pp 7–19

de Soete G, Winsberg S (1993) A thurstonian pairwise choice model with univariate and multivariate spline transformations. Psychometrika 58(2):233–256

Dai Y, Lim TC, Karr CL (2003) Subjective response simulation of brake squeal noise applying neural network approach. Noise Control Eng J 51(1):50–59

Danna J, Velay JL, Paz-Villagran V, Capel A, Petroz C, Gondre C, Thoret E, Aramaki M, Ystad Y, Kronland-Martinet R (2013) Handwriting movement sonification for rehabilitation of dysgraphia. In: Proceedings of the 10th Symposium on Computer Music Multidisciplinary Research, Marseille, 2013

Danna J, Fontaine M, Paz-Villagran V, Gondre C, Thoret E, Aramaki M, Kronland-Martinet R, Ystad S, Velay JL (2015) The effect of real-time auditory feedback on learning new characters. Hum Mov Sci 43:216–228

Ellermeier W, Mader M, Daniel P (2004) Scaling the unpleasantness of sounds according to the BTL model: ratio-scales representation and psychoacoustical analysis. Acta Acust united Ac 90:101–107

Elliott TM, Hamilton LS, Theunissen FE (2013) Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. J Acoust Soc Am 133(1):389–404

Gaver WW (1986) Auditory icons: Using sound in computer interfaces. Hum Comput Interact 2(2):167–177

Gaver WW (1989) The sonic finder: An interface that uses auditory icons. Hum Comput Interact 4:67–94

Grey JM (1977) Multidimensional perceptual scaling of musical timbres. J Acoust Soc Am 61(5):1270–1277

Hansen H, Verhey JL, Weber R (2011) The magnitude of tonal content, a review. Acta Acust united Ac 97(3):355–363

Hoffmann A, Bergman P, Kropp W (2016) Perception of tyre noise: can tyre noise be differentiated and characterized by the perception of a listener outside the car? Acta Acust united Ac 102(6):992–998

International Telecom Union (2001–2003) Recommendation ITU-R BS.1534–1: method for the subjective assessment of intermediate quality level of coding systems. International Telecom Union, Geneva

Jeon JY, You J, Chang HY (2007) Sound radiation and sound quality characteristics of refrigerator noise in real living environments. Appl Acoust 68:1118–1134

Jeong U-C, Kim J-S, Jeong J-E, Yang I-H, Oh J-E (2015) Development of a sound quality index for the wash cycle process of front-loading washing machines considering the impacts of individual noise sources. Appl Acoust 87:183–189

Krumhansl C (1989) Why is musical timbre so hard to understand? In: Nielzen S, Olsson O (eds) Structure and perception of electroacoustic sound and music. Elsevier, Amsterdam, pp 43–53

Lartillot O, Toiviainen P (2007) A Matlab toolbox for musical feature extraction from audio. In: Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07), University of Bordeaux 1, France

Lemaitre G, Susini P, Winsberg S, Letinturier B, McAdams S (2007) The sound quality of car horns: a psychoacoustical study of timbre. Acta Acust united Ac 93(3):457–468

Lemaitre G, Houix O, Visell Y, Franinovic K, Misdariis N, Susini P (2009) Toward the design and evaluation of continuous sound in tangible interfaces: the Spinotron. Int J Hum Comput Stud 67:976–993

Lemaitre G, Vartanian C, Lambourg C, Boussard P (2015a) A psychoacoustical study of wind buffeting noise. Appl Acoust 95:1–12

Lemaitre G, Heller LM, Navolio N, Zuñiga-Peñaranda N (2015b) Priming gestures with sounds. PLoS One 10(11):e0141791

McAdams S, Winsberg S, Donnadieu S, Soete GD, Krimphoff J (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities and latent subject classes. Psychol Res 58:177–192

Parizet E, Brocard J, Piquet B (2004) Influence of noise and vibration to comfort in diesel engine cars running at idle. Acta Acust united Ac 90:987–993

Parizet E, Guyader E, Nosulenko V (2008) Analysis of car door closing sound quality. Appl Acoust 69:12–22

Parizet E, Ellermeier W, Robart R (2014) Auditory warnings for electric vehicles: detectability in normal-vision and visually-impaired listeners. Appl Acoust 86:50–58

Paté A, Lavandier C, Minard A, Le Griffon I (2017) Perceived unpleasantness of aircraft flyover noise: influence of temporal parameters. Acta Acust united Ac 103(1):34–47

Peeters G, Giordano BL, Susini P, Misdariis N, McAdams S (2011) The timbre toolbox: extracting audio descriptors from musical signals. J Acoust Soc Am 130(5):2902–29016

Pietila G, Lim TC (2012) Intelligent systems approaches to product sound quality evaluations – A review. Appl Acoust 73(10):987–1002

Pietila G, Lim TC (2015) Sound quality preference modeling using a nested artificial neural network architecture. Noise Control Eng J 63(2):138–151

Rath M, Schleicher R (2008) On the relevance of auditory feedback for quality of control in a balancing task. Acta Acust united Ac 94(1):12–20

Sato S, You J, Jeon J (2007) Sound quality characteristics of refrigerator noise in real living environments with relation to psychoacoustical and autocorrelation function parameters. J Acoust Soc Am 122(1):314–325

Schaeffer P (1966) Traité des objets musicaux. Editions du Seuil, Paris

Serafin S, Franinovic K, Hermann T, Lemaitre G, Rinott M, Rocchesso D (2011) Sonic interaction design. In: Hermann T, Hunt A, Neuhoff JG (eds) Sonification handbook. Logos Verlag, Berlin, pp 87–110

Siedenburg K, Fujinaga I, McAdams S (2016) A comparison of approaches to timbre descriptors in music information retrieval and music psychology. J New Music Res 45(1):27–41

Stanton NA, Edworthy J (1999) Auditory warnings and displays: an overview. In: Stanton NA, Edworthy J (eds) Human factors in auditory warnings. Ashgate, Aldershot

Susini P, McAdams S, Winsberg S (1999) A multidimensional technique for sound quality assessment. Acta Acust united Ac 85:650–656

Susini P, McAdams S, Winsberg S, Perry S, Vieillard S, Rodet X (2004) Characterizing the sound quality of air-conditioning noise. Appl Acoust 65(8):763–790

Susini P, Lemaitre G, McAdams S (2011) Psychological measurement for sound description and evaluation. In: Berglund B, Rossi GB, Townsend JT, Pendrill LR (eds) Measurement with persons–Theory, methods and implementation area. Psychology Press, Taylor and Francis, New York/London

Susini P, Houix O, Misdariis N (2014) Sound design: an applied, experimental framework to study the perception of everyday sounds. The New Soundtrack 4(2):103–121

Tajadura-Jiménez A, Liu B, Bianchi-Berthouze N, Bevilacqua F (2014) Using sound in multi-touch interfaces to change materiality and touch behavior. In: Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational. Helsinki, Finland, pp 199–202

Terhardt E, Stoll G, Seewann M (1982) Algorithm for extraction of pitch and pitch salience from complex tonal signals. J Acoust Soc Am 71(3):679–688

Thurstone LL (1927) A law of comparative judgment. Psycho Rev 34(4):273–286

Zwicker E, Fastl H (1990) Psychoacoustics facts and models. Springer Verlag, Berlin

Chapter 10
Timbre Perception with Cochlear Implants

Jeremy Marozeau and Wiebke Lamping

Abstract The perception of timbre is fairly well understood for normal-hearing listeners; however, it is still unclear how hearing impairment affects this percept. This chapter addresses how people with severe hearing loss who have been fitted with a cochlear implant perceive timbre. A cochlear implant is a medical device that allows a deaf person to perceive sounds by stimulating their auditory nerve directly. Unlike a pair of glasses that perfectly restores sight, cochlear implants dramatically alter the audio signal. This chapter starts with a brief overview of the design and functioning of a cochlear implant, which is then followed by a discussion of how cochlear implant listeners perceive and identify musical instruments. Thereafter, insights on how cochlear implant listeners perceive the sound quality induced by simple electrical pulse trains will be provided. Finally, the chapter proposes some potential avenues to improve the sound quality experienced through a cochlear implant.

Keywords Deafness · Hearing impaired · Hearing loss · Instrument recognition · Multidimensional scaling · Music perception · Sound quality

10.1 Introduction

In *The Matrix* series of movies (released between 1999 and 2003), the protagonists had a plug on the back of their heads that allowed the direct stimulation of different parts of their brains that included the auditory cortex. With this implant, protagonists were able to hear, see, feel, and taste with such precision that they could not easily distinguish reality from the simulated world. Although those science fiction movies depict a futuristic world, similar technologies have been developed since the

J. Marozeau (✉) · W. Lamping
Hearing Systems Group, Department of Health Technology, Technical University of Denmark, Lyngby, Denmark
e-mail: jemaroz@dtu.dk; wila@dtu.dk

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_10

1970s to enable hearing in some humans who suffer from hearing impairment. This device, the cochlear implant (CI), is a neural prosthesis that can be prescribed for people with severe to profound hearing loss when they are not able to gain much from a traditional hearing aid. The CI can restore the sensation of hearing by electrically stimulating the auditory nerve directly, bypassing the malfunctioning receptor cells. However, unlike the brain implant from "Matrix," the ability of a CI to restore an auditory sensation is limited and can vary substantially across recipients. This chapter will give some insight on how people wearing a CI might perceive timbre.

Assessing sound quality through a CI is not an easy task as we are lacking the vocabulary to describe sounds. For most people, it is challenging to describe the sound of a car in terms of basic auditory elements: "The sound of a car passing sounds like … a car passing". Furthermore, even if someone knows a lot about timbre semantics (Saitis and Weinzierl, Chap. 5), how can we make sure that their "warm" or "bright" sound through a CI relates to a similar sensation in acoustic hearing?

As described by Siedenburg, Saitis, and McAdams (Chap. 1) and in greater detail by McAdams (Chap. 2), many tools have been developed to study the relationship between acoustic properties, such as the frequency or the temporal envelope, and the perceptual dimensions of timbre. If some of those tools are to be adapted for CI listeners, it is essential to adapt them to the CI-specific sound processor. For example, a sizeable acoustic variation can result in an insignificant physical change in the pattern of the electrical stimulation. Therefore, the relationship between acoustic properties and the dimensions of the timbre (such as between the spectral centroid and the brightness) needs to be redefined with the physical parameters that characterize the electrical stimulation. Additionally, it is essential to bear in mind that some of the perceptual dimensions experienced by normal-hearing listeners cannot be translated directly into the perceptual dimensions experienced by CI users.

This chapter first introduces the reader to the technology of the cochlear implant (Sect.10.2). The general limitations of the perception of music through a CI is then discussed (Sect. 10.3) before special attention is dedicated to the identification of instruments (Sect. 10.4). In order to understand how CI users perceive the timbre of musical instruments, studies using the technique of multidimensional scaling are reviewed in Sect. 10.5. Subsequently, Sect. 10.6 focusses on discussing the perception of the most simple stimulus for a CI: the pulse train. Finally, Sect. 10.7 proposes some potential directions on how to improve the perception of timbre in a CI.

10.2 What Is a Cochlear Implant?

A healthy human cochlea contains, on average, 3500 inner hair cells located along the basilar membrane that will provide input to 30,000 auditory neurons that carry auditory information to the brain. Unfortunately, damage to those inner hair cells, for example by exposure to loud sounds, is irreversible. This damage, commonly known as a sensorineural hearing loss, will prevent the acoustic wave from triggering action potentials on the auditory nerve. When the portion of remaining healthy inner hair cells is high, a hearing aid can partially restore hearing by amplifying the sound. However, for more severe damage, even the most powerfull hearing aid will not be able to provide sufficient gain to restore speech intelligibility. A person suffering such profound loss (corresponding to hearing thresholds elevated by at least 80 dB) will be considered a candidate for a CI as this device can directly stimulate the auditory nerve and replace the function of the damaged inner hair cells. The specific criteria for the eligibility for a CI vary across countries, and candidacy is often determined on a case-to-case basis, for which multiple factors, like the auditory profile but also development, cognition, or psychosocial functioning, are of importance (Niparko 2004).

A CI replaces the different roles of the outer, middle, and most of the inner ear. It is composed of one external part and one internal part. First, the sound processor, just like a typical hearing aid, is hooked behind the ear (Fig. 10.1, #1). It contains one or more microphones, batteries, and an electronic circuit that converts the acoustic signal into a digital signal, according to a programmable software algorithm called a *strategy*. This signal is transmitted via a wireless radio-frequency link antenna (Fig. 10.1, #2) across the skin to the implant's receiver (Fig. 10.1, #3), which is aligned to the receiver by a pair of magnets. The radio frequency signal is then decoded into a series of electric pulses. The internal part of the implant includes a linear array of up to twenty-two electrodes inserted about halfway into the spiral-shaped cochlea. These electrodes directly stimulate the auditory nerve, thus replacing the function of the hair cells that are lost or damaged in sensorineural deafness.

There are many parameters, such as electrode location, the number of active electrodes, stimulation rate, phase duration, current amplitude, or polarity, that may influence how an electrical stimulus is being perceived. For the sake of brevity, only some basic parameters that affect loudness and pitch and how a CI transmits sounds

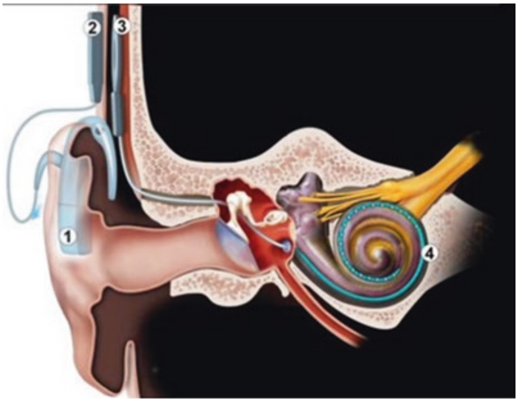

Fig. 10.1 Schematic of a cochlear implant. A sound processor (*1*) captures sounds via a microphone and converts them into an electrical signal. This signal is then transmitted through the skin by a transmitter (*2*) to the internal component composed of a receiver (*3*) and an electrode array (*4*). (Drawing by S. Blatrix for R. Pujol, www.cochlea.eu; used with permission of R. Pujol)
]

to the recipient will be covered. For a more detailed review of psychophysical studies in electric hearing, the reader is referred to McKay (2004) or Zeng et al. (2008).

When an electrode is activated, it delivers a series of biphasic symmetric pulses, generally with phase durations of 25–100 μs and a 0–20 μs interphase gap. The loudness of such a pulse train depends on the overall electric charge but also on other factors, for example, the distance of the electrode to the neurons, the survival of neurons close to the electrode, the rate of stimulation or activation of the other electrodes (McKay et al. 2003; McKay 2004; Bierer 2010).

A change in pitch can be elicited by changing the place of the active electrode. If the electrode is located close to the cochlear apex, the sensation induced is often described as low-pitched by the patient, and if it is closer to the base of the cochlea, it is described as high-pitched, following the tonotopic organization of the auditory nerve (place pitch). Similarly, an increase in pitch can also be achieved by increasing the pulse rate (rate pitch) in pulses per second (pps) on a single electrode (e.g., Zeng 2002).

The strategy embedded in the sound processor determines which combinations of electrodes stimulate the auditory nerve. This is based on an analysis of the acoustic signal received by the microphone. Fig. 10.2 shows a simplified block diagram of how the CI transforms a sound into patterns of electric stimulation. The most commonly used strategy, named the *advanced combination encoder* (ACE), divides the incoming sound signal into as many frequency bands as there are electrodes,

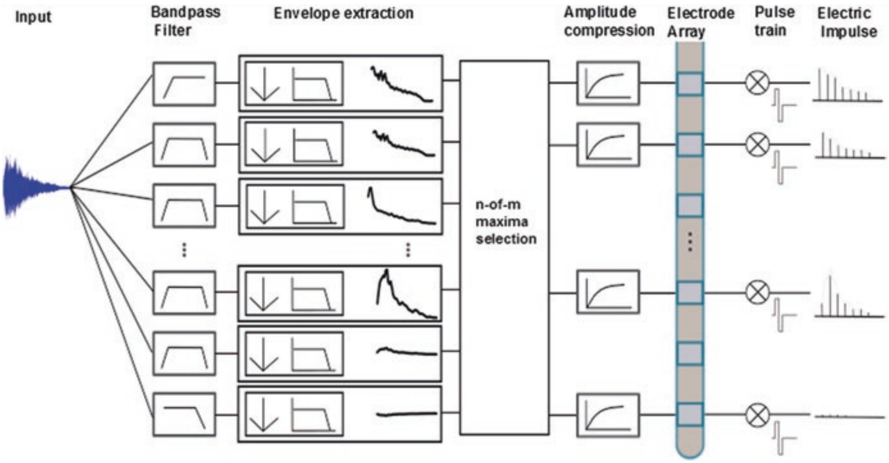

Fig. 10.2 Simplified block diagram of a cochlear implant sound-coding strategy. First the sound is decomposed though a band pass filter. Then the output of each band is processed through an envelope extraction composed of an envelope detection block and a low pass filter. For each time frame and each band, the energy is evaluated and compared. Only a subgroup (n electrodes selected out of the m band) of the bands with the maximum energy will be selected, typically 8 out of 22. A nonlinear compressive function will be applied on the energy in the selected band in order to convert it to a current value in mA. Finally, on the electrodes assigned to the selected bands, a series of electrical impulses will be produced and modulated at a fixed rate with the envelope output of the associate band

selects a subnumber of the bands with the highest amplitude, typically 8–12, and then stimulates the associated electrodes in an interleaved manner at a current level related to the energy within the band (McDermott et al. 1992). On each of the selected electrodes, the electrical pulse train will be produced at a fixed rate, which is typically between 500 and 5000 pps and is modulated with the envelope output of the associated band.

One might wonder why the CI is limited to only a maximum of twenty-two electrodes. To begin with, the cochlea is a tiny structure and to be able to insert that many electrodes and their wires is already a great technological achievement. Furthermore, because of the distance between the electrodes and the auditory neurons, the structure of bony wall that separates the electrode from the nerve, and the high conductivity of the fluid inside the cochlea, the current generated cannot be restricted to the vicinity of the target neurons. That means that each electrode will stimulate a broad range of sensory neurons, and some electrodes may not be discriminable for the listener, that is, two or more electrodes elicit the same or a similar place-pitch sensation. The resulting frequency resolution may be good enough for CI users to perceive sentences in a quiet environment but will prevent a clear perception of more complex sound environments (Blamey et al. 2013). Further, rate-pitch cues are not being incorporated in most of the contemporary processing strategies. Almost all manufacturers, apart from one, exclusively deliver the signal envelope on a fixed-rate carrier. First, rate pitch is limited as it has been shown to reach a plateau, meaning that the pitch saturates after a specific rate or upper limit, which varies across individuals but is roughly around 300 pps (Tong et al. 1983, see also Sect 10.6.2). Second, it is difficult to determine a proper place-rate match (Oxenham et al. 2004) that will indicate which rate should be chosen for which electrode. Here, shallow and variable insertion depths of the electrode array and the broad spread of excitation along the cochlea are restricting the possibility for matching temporal and spectral information thoroughly. For this reason, temporal fine-structure cues, which have been shown to play an important role in pitch and speech perception (Moore 2008), are often not transmitted.

There is considerable variability in the reactions among patients whose CIs are activated for the first time. Some people will regain speech perception right away while others might feel great discomfort and eventually reject the device. For most patients, the brain needs time to adapt to this new type of sensory input, and patients will have to be exposed to the CI for a few months before they can fully benefit from their new device. During this time, timbre and pitch perception will change significantly. Anecdotally, one CI user described that other people’s voices sounded very high pitched, like Alvin the chipmunk (the animated movie character) at activation, and after a few months the perception had deepened to sound like Darth Vader (the Star Wars movie character). Nevertheless, a significant portion of CI users can understand speech correctly in quiet listening situations soon after they get the device. Unfortunately, many users still struggle with more challenging listening environments, such as speech in noise, or they may have difficulties with music appreciation (Zeng et al. 2008).

10.3 The Perception of Music with a Cochlear Implant

Many CI users are now able to understand words and sentences in quiet listening environments without any other aid, such as lip-reading (Blamey et al. 2013). Although understanding speech is most important for the majority of recipients, many CI users have expressed their wish to have their perception of music fully restored. A survey of the musical habits of over hundred CI listeners found that music was generally less enjoyable post-implantation (Looi and She 2010). However, this survey also showed the large variability toward the amount of time spent listening to music and the overall enjoyment of music. Some CI listeners reported that they never listened to music and received no enjoyment at all, while others reported listening to music very often and found great joy in that activity. This latter group can be surprising given the limitations of most CI listeners in the perception of fundamental musical features.

Several studies have shown that the majority of CI users have difficulties with perceiving pitch accurately. Using different tasks, such as pitch discrimination, melody recognition, or pitch change detection, CI users perform, on average, significantly worse than normal-hearing listeners (e.g., Laneau et al. 2006; Kang et al. 2009). In most western music, the semitone is the smallest standard pitch difference between two notes (about 6% difference in fundamental frequency, F_0). However, CI users need an average difference of at least 25% in F_0 between notes, more than a minor third, to start to accurately assess the direction of pitch change (Looi et al. 2004). Given that the most common musical intervals between two notes are below a third (Huron 2001), it is clear how challenging it can be for CI listeners to follow a melody. Fortunately, CI listeners can partly rely on rhythm cues to correctly identify well-known melodies (Kong et al. 2004). This ability allows CI listeners to dance in time to the beat of the music (Phillips-Silver et al. 2015) or correctly identify the intended emotion of a piece of music (Vannson et al. 2015).

Along with the difficulties in pitch discrimination, several studies have reported some challenges in the perception of the timbre of complex sounds (McDermott 2011). One might argue that the perception of timbre is only secondary to the enjoyment of music. Is it really essential to be able to recognize a trumpet from a trombone to appreciate the ninth symphony of Dvorak? Even if the sound quality of a piece of music can be greatly reduced by a cheap sound reproduction system, it is still possible to enjoy the music to a certain extent! While this might be true, timbre is not only necessary to convey subtle musical information, but it is also of paramount importance to segregate different melodic lines. If a listener cannot hear the timbral differences between the instruments of an orchestra, they will lose the ability to segregate the instruments, and a beautiful and sophisticated symphony may then turn into a giant sonic mess.

10.4 Instrument Identification

To test musical instrument identification by CI listeners, Gfeller et al. (2002a) asked fifty-one participants to identify sounds from eight different musical instruments. They compared the scores of CI listeners with scores from twenty normal-hearing listeners. CI listeners showed a mean performance of 47% correct responses, whereas normal-hearing listeners achieved an average score of 91%.

It is possible that musical training can influence the identification score greatly and that CI users, because of their hearing impairment, were generally less musically trained. However, in another study by the same group (Gfeller et al. 2002b), CI listeners were trained for 12 weeks on timbre recognition and appraisal. Listeners were asked to identify sixteen musical instruments before and after the training. A positive effect of training was found as the recognition score increased significantly from 34% to 58% correct. However, even after extensive training, the final score still remained much lower than the one obtained for normal-hearing listeners without training.

A similar experiment was performed by McDermott and Looi (2004), who found comparable results with ten CI listeners. The stimuli were composed of sixteen music samples of individual instruments (played for 5 s). Each sample was representative of the type of music played by this instrument (for example rock music was played on the drums and classical music on the tympani). Therefore, participants were not only able to rely on timbre cues to identify the instrument, but they could also rely on pitch cues and genre-specific information. On average, the normal-hearing listeners were able to identify the musical instrument (97% correct responses) while the CI listeners showed a much lower score (47%).

Figure 10.3 shows the confusion matrix of the CI listeners. The stimuli can be divided into three groups: sustained musical instrument, the human voice, and percussion instruments. As is common in musical instrument identification tasks (see McAdams Chap. 2), most of the confusions occurred within groups. For example, the male voice was often confused with the female voice, the trumpet with the flute, and the piano with the harp. A few exceptions were noted, for example, the female voice was confused with the trumpet and the timpani with the double-bass. This latter percussive instrument can play well-defined melodies by varying the tension of the skin through a pedal. It might then be easily confused with a double bass, which is often played pizzicato in jazz music. Finally, the organ was poorly identified; this might be caused by the lack of familiarity for that specific instrument.

By analyzing the physical features of the instruments, it appears that the temporal envelopes of the stimuli might have played an essential role for identification. Both the piano and the harp are impulsive instruments with sharp attacks and long decays; on the other hand, both the violin and the organ are sustained instruments and have slower attack times. Finally both female and male voices share a similar temporal envelope with a vibrato specific to the human voice. Although the CI listeners have some difficulties in processing spectral information, the temporal envelope is fairly well conveyed. It is worth noting that the samples from the male and female voices were recordings of lyrical singers who sang with a pitch range much

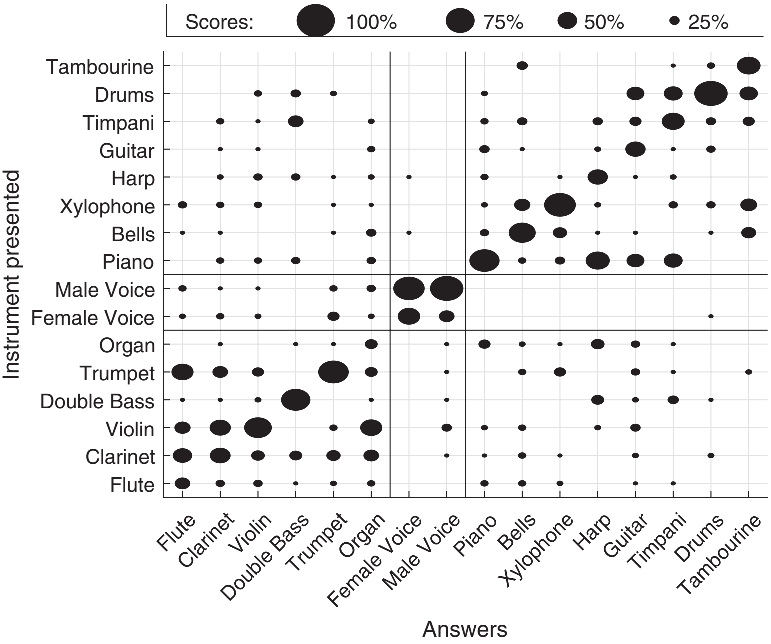

Fig. 10.3 Confusion matrix on an instrument identification task. *Filled circles* indicate relative scores (percent correct) as indicated on the figure. The data have been replotted based on the confusion matrix found in McDermott (2004)

higher than their everyday speaking voices. This might partly explain why the male voice was often confused with the female voice, but this result is also in agreement with various studies showing that gender categorization by CI users is generally poorer than the categorization performance of normal-hearing listeners (Fu et al. 2004, Fu et al. 2005).

Children who have been given an implant as infants display similar behaviors (Innes-Brown et al. 2013). Results show that despite 1 year of musical training, children with a CI still underperformed in instrument recognition tasks compared to their aged-matched normal-hearing peers with the same amount of musical education.

As seen in Sect. 10.2, the cochlear implant divides the overall spectrum in 12 to 22 bands (depending on the manufacturer), and each of those electrodes will activate a broad range of overlapping auditory neurons. Therefore, information related to the spectral envelope will be highly degraded in electric hearing compared to acoustic hearing. On the other hand, the sound processor operates with a reasonably fast time window (at the level of milliseconds), and the overall temporal envelope of the signal can be preserved. Many previous studies on timbre have demonstrated that the temporal envelope is one of the most important factors in the perception of timbre (see McAdams, Chap. 2). From the evidence described above, it might then be hypothesized that CI listeners would rely on the temporal dimension more than on the spectral dimension.

10.5 Multidimensional Scaling Studies

Timbre is a complex percept that can be decomposed into different perceptual dimensions (see McAdams, Chap. 2). The two most important dimensions are the impulsiveness, which correlates with the logarithm of the attack time, and the brightness, which correlates with the spectral centroid. An advanced technique of multidimensional scaling analysis (MDS) allows us to derive the perceptual weight that each listener will assign to those dimensions (see information on CLASCAL in McAdams, Chap. 2). Such weights will depend primarily on the set of stimuli but might also depend on the hearing impairment of the listeners.

In 2012, an experiment by McAdams et al. (1995) was reproduced with seven bimodal (one CI and one hearing aid) and five bilateral (two CIs) listeners (Kong et al. 2011, 2012). The goal of the study was to evaluate the perceptual weight that CI users put on the dimensions correlated to the temporal and spectral descriptors. Kong and colleagues hypothesized that, based on studies on instrument recognition, the temporal envelope cues will dominate, and the reliance on spectral cues will be reduced for bilateral users due to current spread and limited spectral resolution. They further hypothesized that bimodal listeners with some residual low-frequency hearing and with the help of a hearing aid would be able to rely more on the spectral aspect than bilateral users.

Listeners were asked to rate the dissimilarity between all possible pairs of a set of sixteen stimuli. Those stimuli were synthesized western musical instruments similar to those used by McAdams et al. (1995). Each participant was tested in three listening conditions: individual device alone (single CI or hearing aid alone) and with the combined use of devices. The dissimilarity results were transformed into two-dimensional spaces oriented to provide as good a fit as possible with the first two dimensions of the three-dimensional space obtained by McAdams and colleagues. A two-dimensional solution was preferred to a three-dimensional solution, as the model with three dimensions did not improve the goodness-of-fit significantly and could not be interpreted.

Different factors might explain the lack of a third dimension. First, the number of participants was much lower in this study (n = 12 CI listeners) compared to the study of McAdams and colleagues with 98 normal-hearing listeners.[1] The lower the number of participants, the less stable the solution will be; thus this factor will directly impact the number of dimensions that can be extracted. Second, the number of stimuli was slightly lower, 18 in McAdams et al. (1995) versus 16 in the studies by Kong et al. (2011, 2012) to keep the duration of the experiment reasonable while maintaining the appropriate number of repetitions. The number of stimuli can also impact the maximal number of dimensions that can be represented reliability in MDS. Finally, it might be possible that CI users cannot perceive the third dimension consistently. For normal-hearing listeners, this dimension was correlated with sub-

[1] It is worth noting that in CI studies, the number of subjects is often very small as it is quite challenging to find suitable participants.

tle spectrotemporal variation cues, known as the *spectral flux*. It separates stimuli with a fluctuating spectrum, like the vibraphone, from stimuli with a steady spectral envelope, like the bassoon. A possible explanation for this may be that the sound processor would not transmit these cues or that CI users are not able to perceive them.

Each of the two dimensions found with the CI users was correlated with the descriptors known to correlate well with the perceptual dimensions obtained with normal-hearing listeners (Peeters et al. 2011). The left panel of Fig. 10.4 shows that the first dimension was highly correlated with a measure of impulsiveness (as described in Marozeau et al. 2003). This correlation was highly significant for the three conditions of both groups (bilateral and bimodal). As a similar level of correlation is often found with spaces extracted with normal-hearing listeners (see McAdams, Chap. 2), this result suggests that CI users (bimodal and bilateral) perceive the first dimension of timbre similarly compared to the normal-hearing listeners. This assumption is further supported by the results of the timbre recognition task, described previously, in which CI listeners seem to rely mostly on the temporal envelope for stimulus identification.

The second dimension is often found to be highly correlated with the spectral centroid (see Fig. 10.4, *right panel*). However, for the CI listeners, the second dimension was only moderately correlated with the spectral centroid, and no other descriptor with a better correlation was found. The weak correlation with the spectral centroid stands in contrast to results from the normal-hearing listeners (McAdams et al. 1995). It is unclear whether CI users rely on a different physical

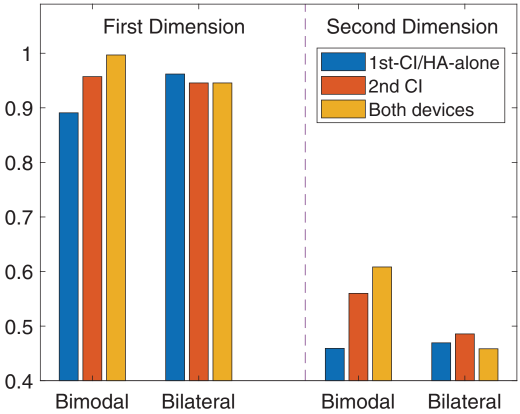

Fig. 10.4 Correlations between perceptual dimensions and physical descriptors of timbre. The *left panel* shows the correlations between impulsiveness and the projection on the first dimension at the group level for the bimodal (one CI or one HA) and the bilateral groups (two CIs) for the three conditions: individual device alone (single CI or hearing aid alone) and with the combined use of devices. This correlation is very high for the normal-hearing listeners. The *right panel* shows the correlations between the spectral centroid and the projection on the second dimension at the group level for the bimodal group and the bilateral group for the three conditions. *CI*, cochlear implant; *HA*, hearing aid (Data replotted from Kong et al. 2012)

cue or whether they rely on the same dimension as the normal-hearing listeners but with a weaker perceptual resolution.

The weight that each listener had put on the two dimensions was evaluated by using an INDSCAL analysis (see McAdams, Chap. 2). The ratios of the weights between the two dimensions can indicate if listeners relied more on the temporal dimension compared to the spectral dimension. This analysis showed that both bilateral and bimodal listeners put more weight on the first dimension compared to the second dimension. For the bimodal listeners, the listening conditions also had an effect: the contribution of the spectral dimension was larger in the condition with a CI (CI-alone and CI plus hearing aid) compared to the condition with hearing aid alone.

These results might seem surprising; however, it is worth noting that bimodal users have very limited residual hearing, mostly in the low-frequency range. As noted above, to be a candidate for a CI, the recipient needs to demonstrate that even with a hearing aid, they cannot understand speech at a functional level. Therefore, if bimodal patients cannot perceive different vowels with their hearing aid, it is unlikely that they can use spectral cues to perceive timbre. All the sounds will be rather dull, and listeners might have to rely mainly on temporal information. If they had normal hearing on one side (so-called single-sided deafness), it is possible that they would rely more on spectral cues through their acoustic hearing. It is also interesting to note that in the CI-alone condition the participants relied more on the second dimension compared to bimodal listeners. This observation might also be supported by the significant correlation between the second dimension and the spectral centroid for the CI-alone condition compared to the condition with the hearing aid alone.

In summary, CI users seem to use the same cues and put similar weight as normal-hearing listeners on the first dimension, which is significantly correlated with the impulsiveness of a sound. The second dimension is less clear and may be related to the spectral centroid, but more studies need to explore the perception of this spectral dimension.

10.6 The Perception of a Single Pulse Train

To be able to understand how CI users perceive the timbre of a complex sound, it is useful to analyze their auditory perception of simple elements. Just as in acoustic hearing when it is useful to consider sounds as a superposition of pure tones, in electric hearing it is useful to consider sounds as sums of pulse trains. This section will discuss and summarize different experiments that study how CI users perceive the sound quality of single-electrode pulse trains when bypassing the signal processor to control the exact pattern of stimulation. Three of the most popular methods that have been used to study timbre of musical sounds will be reviewed in the framework of the perception of electrical pulse trains: (1) MDS analyses, (2) verbal attribute magnitude estimation (VAME), and (3) timbre production.

10.6.1 Multidimensional Scaling Analysis

The pulse train is a series of symmetric biphasic electric current pulses regularly spaced in time. Pulses are composed of two identical phases of opposite polarity to avoid charge accumulation inside the cochlea. As mentioned in Sect. 10.2, the overall charge is associated with the dimension of loudness (i.e., low-charge pulses are perceived as soft and high-charge pulses are perceived as loud); the repetition rate of the pulses is associated with the dimension of pitch, often referred to as the *rate pitch cue*. Electrodes deeply inserted into the cochlea, toward the apex, will induce a lower pitch sensation than the electrodes located near the base, referred to as the *place pitch cue*. Zeng (2002) confirmed the assumptions by asking CI users to rate the pitch height of different pulse trains varying in rate and electrode position. They showed that a high-rate pulse train was perceived as higher than a low-rate and a pulse train delivered on an electrode placed closer to the apex was perceived as lower than on an electrode located at the base. Therefore, it might be tempting to assume that both rate and place pitch cues can be associated with two orthogonal dimensions of pitch. However, McDermott and McKay (1998) argued that the sensation induced by the place cue might be more related to timbre than pitch. If this is the case, the sensation induced by a change of electrode position should be independent to a sensation induced by a rate change.

To test that hypothesis, Tong et al. (1983) asked CI listeners to rate the dissimilarities between nine pulse trains composed of the combination of three rates and three electrode positions. The obtained dissimilarities were represented in a two-dimensional space (see Fig. 10.5). The projection of the stimuli on the first dimension is clearly ordered by pulse rate, while the projection on the second is ordered by electrode placement. This indicates that the two cues (rate and place) will affect two different and orthogonal perceptual dimensions. However, it is unclear how those dimensions are related to the perceptual dimensions experienced by normal-hearing listeners. As a change in electrode placement and a change in rate of stimulation can be ranked on a scale from low to high, one might associate those two dimensions with two dimensions of pitch, such as chroma and tone height. However, as the timbre attribute “brightness” can be scaled from low to high and is highly correlated to the spectral centroid, this second dimension also has been associated with timbre. Even though we often refer to place cues as place “pitch”, it is likely to be the timbre of a sound that changes when different electrodes are stimulated. Overall, the most important conclusion from this study is that the sensation induced by changes in place and rate are not equivalent as they are scaled on different perceptual dimensions.

10.6.2 Verbal Attribute Magnitude Estimation

To understand how CI recipients perceive the sound quality of a single pulse train, it might seem natural just to ask them to describe it. Unfortunately, humans are not trained to talk about sounds. If the recipients describe their sensation freely, it is

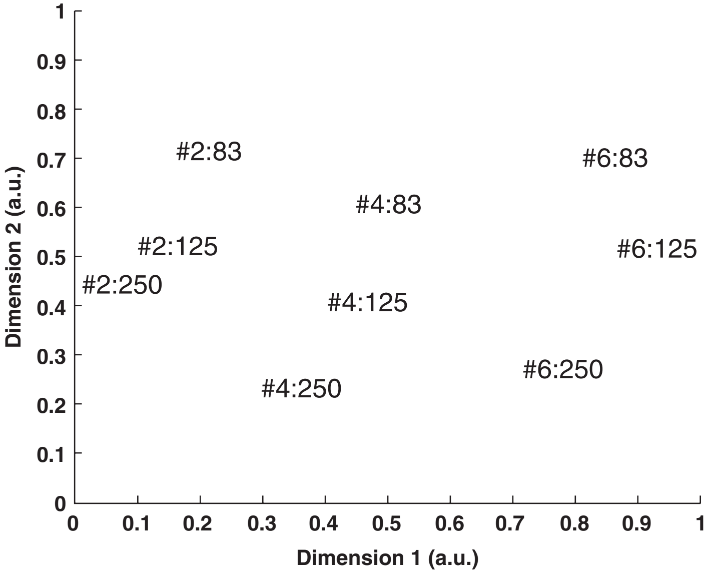

Fig. 10.5 A multidimensional scaling (MDS) solution for responses in the cochlea to pulsed stimuli. The first number indicates the electrode position (*#6 most apical*, *#2 most basal)*, and the second number is the rate in pulses/second. The MDS solution is reproduced in a space with arbitrary units (*a.u.*). (Replotted from Tong et al. 1983)

unlikely that each listener will use the same vocabulary. As a result, it will be very challenging to draw general conclusions from their responses. To avoid this difficulty, von Bismarck (1974) proposed a method based on semantic differentials and factor analysis to extract the common dimensions of timbre that can be described in terms of verbal attributes (also see Saitis and Weinzierl, Chap. 5). In this method, participants are asked to rate stimuli on scales of opposite attributes. Because each pair of attributes is clearly defined, the variability of interpretation for each participant can be significantly reduced. Further, to not expose the listeners to too many different words, Kendall and Carterette (1993a, b) proposed to use only the attribute and its negation (for instance, bright and not bright) instead of using attribute opposites (bright and dull).

Recently, two studies have explored the perceptual qualities of single-electrode pulse trains using this method (Landsberger et al. 2016; Lamping et al. 2018). In the latter one, Lamping and colleagues have asked CI users to rate the pitch, and sound quality of a stimulus set that varied in pulse rate or (sinusoidal) amplitude modulation frequency and electrode placement using multiple verbal attributes. The results indicated that both temporal and place cues could be linked to pitch and timbre.

Figure 10.6A (left) shows that average ratings on the pitch height scale were in agreement with previous findings that showed a significant dependency of pitch on electrode location and pulse rate (Fearn and Wolfe 2000) while showing no interaction between the two of them. As noted above, pitch reaches a plateau at around 300 pps, which can be seen in this study as well as in others (e.g., Zeng 2002; Kong et al. 2009). Most attributes were highly correlated and can be summarized by looking at only two of them: roughness and brightness (Fig. 10.6, middle and right).

Roughness was mainly influenced by the pulse rate. For the region between 80 and 300 pps, roughness decreased monotonically; thereafter, pulse trains were not perceived as rough anymore. The position of the stimulated electrode only had a small effect on roughness with pulse trains being rated as less rough at the apex relative to other locations along the cochlea.

For brightness, both pulse rate and electrode position had a significant effect on the scaling: brightness increased to about 600 pps with increasing pulse rate. For higher rates, there was no effect of rate cue, and place cues completely dominated the perceived brightness. Interestingly, brightness was the only attribute for which there was a significant interaction between the two main factors, suggesting that the two cues may not be entirely independent, at least in their perceived brightness.

Figure 10.6B shows the ratings for modulated pulse trains. Both rate and electrode location showed a significant effect on all three attributes. However, for the brightness attribute, only the rating on electrode 22 (most apical) differed from those for other electrodes. The rating for regular pulse trains and modulated pulse

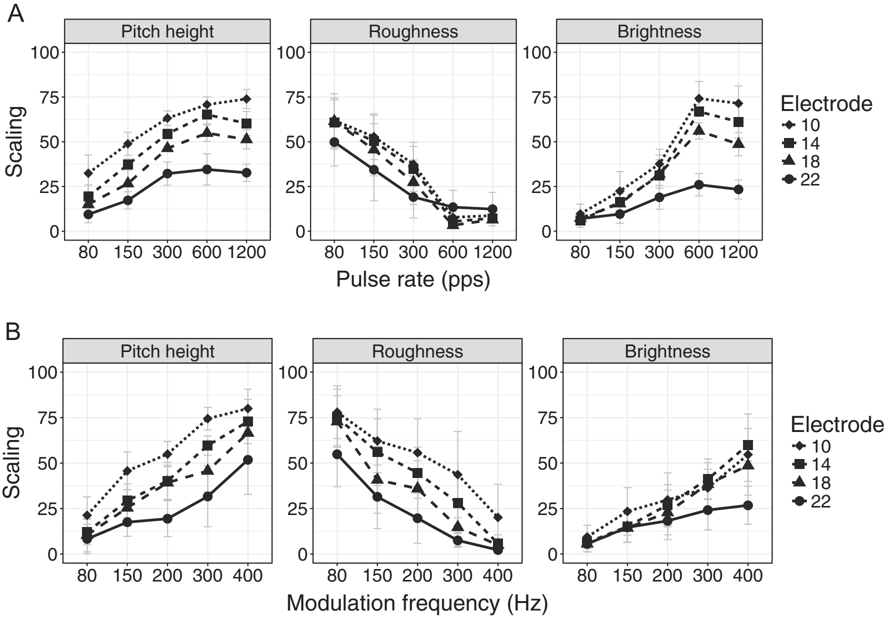

Fig. 10.6 Adjective rating of electrical pulse trains. The upper panels (**A**) show the average of scaled values for all participants for pitch height (*left*), roughness (*middle*), and brightness (*right*) for unmodulated pulse trains. *Electrode 22* is the most apical electrode and *electrode 10* is in the middle of the array (most basal tested). The lower panels (**B**) show the results for the modulated pulse trains on a constant carrier rate of 1200 pps. Error bars depict the standard error. (Data replotted from Lamping et al. 2018)

trains showed reasonably similar trends. There was no significant difference in rating between the results of these two sets for frequencies of 80 Hz, 150 Hz, and 300 Hz. These similarities in sound quality seem consistent with measures of temporal acuity in CI listeners. Kong et al. (2009) showed that rate discrimination thresholds have similar patterns for both modulated and unmodulated pulse trains, potentially indicating a similar pitch salience for these stimuli. The only attribute for which a significant difference between stimulus sets emerged was for roughness, possibly because the corresponding temporal modulation was more explicitly coded by the CI, that is, the remaining pulses between the peaks and troughs of the waveform were still affecting perception.

10.6.3 Timbre Production

As mentioned earlier (Sect. 10.6.2) and discussed by Saitis and Weinzierl (Chap. 5), it can be quite challenging to describe a sound because we lack sensory vocabulary for auditory experiences. Furthermore, it is impossible to know if someone's "red" is similar to someone else's "red"; the degree to which vocabulary is shared across individuals remains unclear, in particular when it comes to CI listeners, who exhibit large interindividual differences. To continue the visual analogy, imagine that you need to describe a sophisticated abstract piece of art such as a painting from Pablo Picasso. You can try to describe an ensemble of round and triangular shapes and dark lines; however, it is hard to convey the essence of the painting. It is much easier to draw a simplified sketch. A similar approach was performed with a specific subgroup of CI users that have near normal hearing in one ear and a CI for the other ear. Recently, CI surgeries have been approved for people with only one impaired ear and disabling tinnitus. The neural mechanisms underlying tinnitus are still relatively unknown, and current theurapeutical approaches mostly rely on psychological intervention. In those patients, the CI was introduced to try to alleviate their tinnitus through auditory input. This new type of patient can provide the researchers with precious information on electric hearing through a CI because it can directly be compared to normal acoustic hearing in the same person.

In this study, 10 CI users were asked to vary the parameters of an acoustic sound played to their normal-hearing ear and to match its perception with that of the electric sensation of two electrodes (Marozeau et al. 2017). Two electrical stimuli were tested: a pulse train on the most apical electrode and a similar pulse train on a mid-array electrode. The participants used a graphical tablet to modify four acoustic parameters of a complex sound. The first two parameters, p1 and p2, were linked to the tonal aspect of the sound and set the frequency of each component according to the formula:

$$\mathrm{fn} = \mathrm{p1}^{*}\mathrm{n}^{\mathrm{p2}}$$

where fn is the frequency of the component n. The other two paramters, p3 and p4, were linked to the timbre of the sound and set the amplitude of those components. The parameter p3 was linked to the spectral centroid and p4 to the spectral spread.

The results showed some substantial differences among the participants, typically observed with CI listeners but possibly enhanced by the complicated task. Among the four test parameters, participants only significantly varied the parameter p3 (spectral centroid) as a function of the electrodes. This can be interpreted as the mid-array electrode inducing a significantly brighter timbre than the apical electrode. On the other hand, the other three parameters (p1, p2, and p4) were not significantly affected by a change of electrode position. The experiment shows once again that the electrode position probably influences timbre rather than the tonality, while changes in pitch might be coded through differences in pulse rate. As most contemporary processing strategies only incorporate a fixed rate parameter, it is easy to comprehend the difficulties that CI users might experience in a task involving the perception of polyphonic music in which they need to extract the pitch of many different instruments playing simultaneously.

10.7 How to Improve the Perception of Timbre

Although cochlear implant technology has existed for over 30 years and CIs have been implanted in more than 600,000 patients around the globe, the overall performance on music perception tasks, or more specifically, timbre-related tasks such as instrument identification (Looi et al. 2012), has not improved substantially for the past 20 years and suffers from several persistent shortcomings. In most modern technologies, hardware and software evolve in parallel. For example, when an update for a computer operating system is released, it is designed for the latest processor technology and would not always be compatible with older generations of computers. On the other hand, in CI technology, once a patient is implanted with a device, they are supposed to keep it for a lifetime. Therefore, although research on new types of implants (CI hardware) is fundamental, many studies aim to improve the outcome of the CI sound processor algorithms that should be compatible with the first generation of implants released more than 30 years ago.

10.7.1 New Sound Processing Strategies

As mentioned in Sect. 10.2, the main limitation in CIs is the spread of current induced by each electrode. Because the inner parts of the cochlea constitute a complex shape with highly conductive fluid (the perilymph), it is challenging to predict where and how exactly the induced current will flow. This complexity will increase if many electrodes are activated simultaneously, so cochlear implant strategies activate at most two electrodes simultaneously. In the most common stimulation mode, called the *monopolar mode*, the current flows between an electrode located inside the cochlea and a return electrode located outside of the cochlea near the temporal bone.

Van den Honert and Kelsall (2007) hypothesized that if the impedance between all electrodes can be measured, it would be possible to create a focused electrical field centered around one specific electrode that should consequently activate narrower regions of auditory neurons. This focused electrical field will be created by selecting simultaneous current levels and phases on each electrode appropriately. Thus the sum of all potentials will result in a narrower field. Marozeau and McKay (2016) have tested this new stimulation mode, called the *all-polar mode*, with five participants and compared the sound quality of this stimulation mode to the monopolar mode. The participants were asked to judge the sound dissimilarity between pairs of stimuli that were presented in either monopolar or all-polar mode. The stimuli were designed to produce two peaks of energy in front of two specific electrodes in order to mimic a vowel with two formants. In the monopolar mode, the two peaks were produced by activating two electrodes sequentially. In the all-polar mode, the currents and phases of all the electrodes were set to produce two separate current fields at the same locations as in the monopolar mode but more focused. Using an MDS technique, a three-dimensional perceptual space was produced (see Fig. 10.7A, B).

The first perceptual dimension was highly correlated with the average position of the two peaks and the second dimension was moderately correlated with the distance between them. Although the electrical field and the spectrum of a sound are two different entities, it is worth noting the similarity between the first correlate of

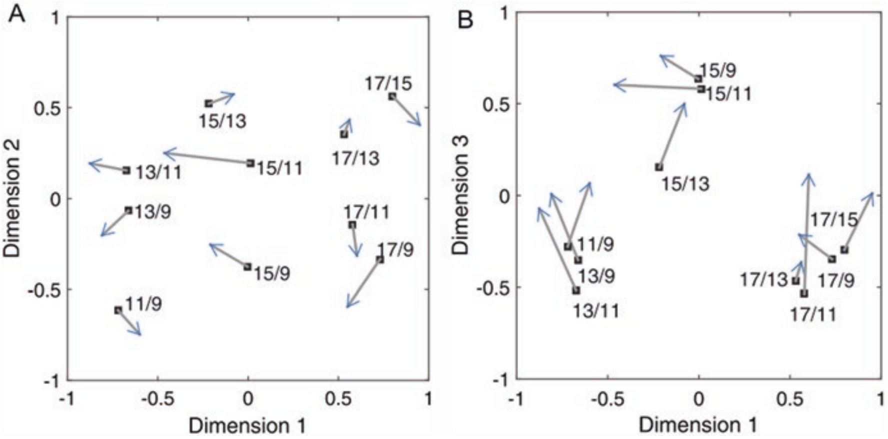

Fig. 10.7 Multidimensional scaling (MDS). (**A**) The first two dimensions of the three-dimensional solution; (**B**) dimension 1 versus dimension 3. Each monopolar stimulus is represented by a *square* and each all-polar stimulus is represented by the end of the *arrow*. The *two numbers* next to each stimulus indicate the location of each formant based on the electrode position within the array. The monopolar and all-polar stimuli that shared the same formant are linked by an arrow. (Data replotted from Marozeau and McKay 2016)

this space (the average location of electrical energy) and the spectral centroid (average frequency of the energy spectrum) and the similarity between the second correlate (the spread of electrical energy) and the spectral spread. More interestingly, the monopolar and all-polar stimuli were significantly separated by a third dimension, which may indicate that all-polar stimuli have a perceptual quality that differs from monopolar stimuli. This experiment suggests that the stimulation mode can have a direct effect on the timbre, more specifically, on the third dimension. It is still unclear how this third dimension can be described. However, using verbal descriptors (clean/dirty) with a similar focused stimulation mode (called *tripolar stimulation*), Padilla and Landsberger (2016) showed that "cleanness" correlated with a reduced spread of excitation. As all-polar mode is supposed to reduce current interaction, it is possible that the third dimension shown in Fig. 10.7 is also related to a "cleaner" sound perception. To summarize, use of these focused stimulation modes, such as all-polar or tripolar stimulation, bears the potential to reduce spectral smearing and thereby improve perceived sound quality.

10.7.2 Training

As noted above, when the CI of a patient is activated for the first time, speech can be hard to understand and the sound sensation may even be confusing. The user will have to go through a hearing rehabilitation program that aims to restore the perception of speech and, sometimes, music and environmental sounds (Philips et al. 2012). However, to restore the perception of timbre, a dedicated training program should be designed. In a survey of eighty-four CI users on what they expected from musical training, Looi and She (2010) found that the ability to recognize commonly known musical instruments was ranked as one of the most desired musical skills that they wished to reacquire, immediately after the ability to recognize known tunes.

Few studies have been developed to target timbre perception specifically. As mentioned in Sect. 10.4, Gfeller et al. (2002b) have found a significant and positive effect of their training protocols. This intensive training program was delivered via a laptop and was composed of 48 lessons over a period of 12 weeks. In each lesson, the student learned about the four principal families of musical instruments (strings, woodwinds, brass, and pitched percussion). Then each instrument was introduced, and information was provided about the material, how it is played, and the typical styles. The students were also gradually exposed to the different sounds that each instrument can produce, starting with a simple note to more complex phrases. Interestingly, not only has this training program resulted in a significant improvement in the task of instrument recognition, but it also had a positive and significant effect on the appraisal of sounds. These results suggest that training can not only help CI users to perceive timbre better but could also enhance the enjoyment of listening to music.

10.8 Conclusions

In summary, users of a cochlear implant still struggle substantially with music appreciation and timbre-related tasks such as musical instrument identification. This may be related to the limited amount of information transmitted via the implant. Even though temporal cues, which are correlated with impulsiveness, receive a similar perceptual weight as in normal-hearing listeners, a degraded spectral envelope emerges due to the limited number of electrodes and current spread. This spread causes overlapping neural populations to be activated when stimulating different electrode locations in the cochlea. Studies that have bypassed the sound processor have shown that the electrode position may be a particularly important cue for timbre perception as it seems to be linked to the timbral attribute of brightness. Finally, a processing strategy transmitting spectral cues more delicately may improve timbre perception, and training to use the new sensory input more efficiently may help with timbre perception.

This chapter has summarized the current experience of CI users and has provided possible solutions using signal processing to improve their perception of timbre. However, it is worth mentioning that many groups around the world are working on solutions for more efficient electrode arrays using light to selectively activate the auditory nerves (Jeschke and Moser 2015) or using gene therapy for the conservation and regeneration of spiral ganglion cells, which would improve the electrode-neural interface (Pfingst et al. 2017). Combining both better selectivity and a better neural interface should allow a major improvement in sound quality and might restore the beautifully complex perception of timbre.

Compliance with Ethics Requirements Jeremy Marozeau declares that he has no conflict of interest.

Wiebke Lamping declares that she has no conflict of interest.

References

Bierer JA (2010) Probing the electrode-neuron interface with focused cochlear implant stimulation. Trends Amplif 14(2):84–95. https://doi.org/10.1177/1084713810375249

Blamey P, Artieres F et al (2013) Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: an update with 2251 patients. Audiol Neurootol 18(1):36–47. https://doi.org/10.1159/000343189

Fearn R, Wolfe J (2000) Relative importance of rate and place: experiments using pitch scaling techniques with cochlear implants recipients. Ann Otol Rhinol Laryngol Suppl 185:51–53

Fu QJ, Chinchilla S et al (2004) The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users. J Assoc Res Otolaryngol 5(3):253–260

Fu QJ, Chinchilla S et al (2005) Voice gender identification by cochlear implant users: the role of spectral and temporal resolution. J Acoust Soc Am 118:1711–1718

Gfeller K, Witt S et al (2002a) Effects of frequency, instrumental family, and Cochlear implant type on timbre recognition and appraisal. Ann Otol Rhinol Laryngol 111(4):349–356. https://doi.org/10.1177/000348940211100412

Gfeller K, Witt S et al (2002b) Effects of training on timbre recognition and appraisal by postlingually deafened cochlear implant recipients. J Am Acad of Audiol 13(3):132–145

Huron D (2001) Tone and voice: a derivation of the rules of voice-leading from perceptual principles. Music Percept 19(1):1–64

Innes-Brown H, Marozeau J et al (2013) Tone, rhythm, and timbre perception in school-age children using cochlear implants and hearing aids. J Am Acad of Audiol 24(9):789–806. https://doi.org/10.3766/jaaa.24.9.4

Jeschke M, Moser T (2015) Considering optogenetic stimulation for cochlear implants. Hear Res (34):322–224. https://doi.org/10.1016/j.heares.2015.01.005

Kang R, Nimmons G et al (2009) Development and validation of the University of Washington Clinical Assessment of music perception test. Ear Hear 30(4):411–418. https://doi.org/10.1097/AUD.0b013e3181a61bc0

Kendall RA, Carterette EC (1993a) Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. Music Percept 10(4):445–467. https://doi.org/10.2307/40285584

Kendall RA, Carterette EC (1993b) Verbal attributes of simultaneous wind instrument timbres: II. Adjectives induced from Piston's "orchestration". Music Percept. https://doi.org/10.2307/40285584

Kong YY, Cruz R et al (2004) Music perception with temporal cues in acoustic and electric hearing. Ear Hear 25(2):173–185

Kong YY, Deeks JM et al (2009) Limits of temporal pitch in cochlear implants. J Acoust Soc Am 125(3):1649–1657. https://doi.org/10.1121/1.3068457

Kong YY, Mullangi A et al (2011) Temporal and spectral cues for musical timbre perception in electric hearing. J Speech Lang Hear Res 54(3):981–994. https://doi.org/10.1044/1092-4388(2010/10-0196

Kong YY, Mullangi A et al (2012) Timbre and speech perception in bimodal and bilateral cochlear-implant listeners. Ear Hear 33(5):645–659. https://doi.org/10.1097/AUD.0b013e318252caae

Lamping W, Santurette S et al (2018) Verbal attribute magnitude estimates of pulse trains across electrode places and stimulation rates in cochlear implant listeners. Proceedings of the international symposium on auditory and audiological research. vol 6 (2017): Adaptive Processes in Hearing

Landsberger D, Vermeire K et al (2016) Qualities of single electrode stimulation as a function of rate and place of stimulation with a cochlear implant. Ear Hear 37(3):e149–e159

Laneau J, Wouters J et al (2006) Improved music perception with explicit pitch coding in cochlear implants. Audiol Neurootol 11(1):38–52

Looi V, She J (2010) Music perception of cochlear implant users: a questionnaire, and its implications for a music training program. Int J Audiol 49(2):116–128

Looi V, McDermott H et al (2004) Pitch discrimination and melody recognition by cochlear implant users. Int Congr Ser 1273:197–200

Looi V, Gfeller K et al (2012) Music appreciation and training for cochlear implant recipients: a review. Semin Hear 33(4):307–334

Marozeau J, McKay CM (2016) Perceptual spaces induced by cochlear implant all-polar stimulation mode. Trends Hear 20. https://doi.org/10.1177/2331216516659251

Marozeau J, de Cheveigné A et al (2003) The dependency of timbre on fundamental frequency. J Acoust Soc Am 114(5):2946–2957

Marozeau J, Ardoint M et al (2017) Acoustic match to electric pulse trains in single-sided deafness cochlear implant recipients. In: International symposium on auditory and audiological research, vol 6, pp 239–246. Danavox Jubilee Foundation. Retrieved from https://proceedings.isaar.eu/index.php/isaarproc/article/view/2017-29

McAdams S, Winsberg S et al (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res 58(3):177–192

McDermot HJ, Looi V (2004) Perception of complex signals, including musical sounds, with cochlear implants. Int Congr Ser 1273:201–204

McDermott HJ (2004) Music perception with cochlear implants: a review. Trends Amplif 8(2):49–82

McDermott HJ (2011) Music perception. In: Zeng F-G, Popper AN, Fay RR (eds) Auditory prostheses: new horizons. Springer, New York, pp 305–339

McDermott HJ, McKay CM (1998) Musical pitch perception with electrical stimulation of the cochlea. J Acoust Soc Am 101(3):1622. https://doi.org/10.1121/1.418177

McDermott HJ, McKay CM et al (1992) A new portable sound processor for the University of Melbourne/nucleus limited multielectrode cochlear implant. J Acoust Soc Am 91(6):3367–3371

McKay CM (2004) Chapter 7: Psychophysics and Electrical Stimulation. In: Zeng F-G, Popper AN, Fay, RR (eds) Cochlear Implants: auditory Prostheses and Electric Hearing. Springer, New York, pp 286–333

McKay CM, Henshall KR et al (2003) A practical method of predicting the loudness of complex electrical stimuli. J Acoust Soc Am 113(4. Pt 1):2054–2063

Moore BCJ (2008) The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. J Ass Res Otolaryng 9(4):399–406. https://doi.org/10.1007/s10162-008-0143-x

Niparko JK (2004) Chapter 3: Cochlear implants: clinical applications. In: Zeng F-G (ed) Cochlear implants: auditory prostheses and electric hearing. Springer, New York, pp 53–100

Oxenham AJ, Bernstein JG et al (2004) Correct tonotopic representation is necessary for complex pitch perception. Proc Natl Acad Sci U S A 101(5):1421–1425

Padilla M, Landsberger DM (2016) Reduction in spread of excitation from current focusing at multiple cochlear locations in cochlear implant users. Hear Res 333:98–107. https://doi.org/10.1016/j.heares.2016.01.002

Peeters G, Giordano BL et al (2011) The timbre toolbox: extracting audio descriptors from musical signals. J Acoust Soc Am 130(5):2902–2916. https://doi.org/10.1121/1.3642604

Pfingst BE, Colesa DJ et al (2017) Neurotrophin gene therapy in deafened ears with Cochlear implants: long-term effects on nerve survival and functional measures. J Assoc Res Otolaryngol 18:731–750. https://doi.org/10.1007/s10162-017-0633-9

Philips B, Vinck B et al (2012) Characteristics and determinants of music appreciation in adult CI users. Eur Arch Otorhinolaryngol 269(3):813–821. https://doi.org/10.1007/s00405-011-1718-4

Phillips-Silver J, Toiviainen P et al (2015) Cochlear implant users move in time to the beat of drum music. Hear Res 321:25–34. https://doi.org/10.1016/j.heares.2014.12.007

Tong YC, Blamey PJ et al (1983) Psychophysical studies evaluating the feasibility of a speech processing strategy for a multiple-channel cochlear implant. J Acoust Soc Am 74(1):73–80

van den Honert C, Kelsall DC (2007) Focused intracochlear electric stimulation with phased array channels. J Acoust Soc Am 121(6):3703–3716. https://doi.org/10.1121/1.2722047

von Bismarck G (1974) Timbre of steady sounds: a factorial investigation of its verbal attributes. Acta Acust United with Acustica 3(14):146–159

Vannson N, Innes-Brown H et al (2015) Dichotic listening can improve perceived clarity of music in cochlear implant users. Trends Hear 19:2331216515598971

Zeng F-G (2002) Temporal pitch in electric hearing. Hear Res 174(1–2):101–106

Zeng F-G, Rebscher S et al (2008) Cochlear implants: system design, integration, and evaluation. IEEE Rev Biomed Eng 1:115–142. https://doi.org/10.1109/RBME.2008.2008250

Part III
Acoustical Modeling

Chapter 11
Audio Content Descriptors of Timbre

Marcelo Caetano, Charalampos Saitis, and Kai Siedenburg

Abstract This chapter introduces acoustic modeling of timbre with the audio descriptors commonly used in music, speech, and environmental sound studies. These descriptors derive from different representations of sound, ranging from the waveform to sophisticated time-frequency transforms. Each representation is more appropriate for a specific aspect of sound description that is dependent on the information captured. Auditory models of both temporal and spectral information can be related to aspects of timbre perception, whereas the excitation-filter model of sound production provides links to the acoustics of sound production. A brief review of the most common representations of audio signals used to extract audio descriptors related to timbre is followed by a discussion of the audio descriptor extraction process using those representations. This chapter covers traditional temporal and spectral descriptors, including harmonic description, time-varying descriptors, and techniques for descriptor selection and descriptor decomposition. The discussion is focused on conceptual aspects of the acoustic modeling of timbre and the relationship between the descriptors and timbre perception, semantics, and cognition, including illustrative examples. The applications covered in this chapter range from timbre psychoacoustics and multimedia descriptions to computer-aided orchestra-

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-3-030-14832-4_11) contains supplementary material, which is available to authorized users.

M. Caetano (✉)
Sound and Music Computing Group, INESC TEC, Porto, Portugal
e-mail: mcaetano@inesctec.pt

C. Saitis
Audio Communication Group, Technische Universität Berlin, Berlin, Germany
e-mail: charalampos.saitis@campus.tu-berlin.de

K. Siedenburg
Department of Medical Physics and Acoustics, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany
e-mail: kai.siedenburg@uni-oldenburg.de

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_11

tion and sound morphing. Finally, the chapter concludes with speculation on the role of deep learning in the future of timbre description and on the challenges of audio content descriptors of timbre.

Keywords Environmental sound · Excitation-filter model · Machine learning · Musical instrument · Pattern recognition · Sound color · Speech · Time-frequency analysis

11.1 Introduction

A sound wave carries a pattern of oscillations, which was generated by a driving force that excited a vibrating object through a physical medium such as the air. When the sound wave reaches the ear, these oscillations are processed and interpreted by the brain as sound. On its way from the source to the ear, the sound wave carries precise information about the vibrating object (e.g., a cello), the driving force (e.g., bowed), and possibly other physical objects with which it interacted (e.g., the walls of a concert hall). The human brain has a remarkable ability to convert the detailed information contained in sound waves into the meaningful experience of hearing—from the minute inflections of speech that facilitate human communication to the expressiveness of microvariations in music (Handel 1995). But how do sound waves convey identifiable properties of the sound source, of the sound-generating event, and even of the objects with which the sound wave interacted? What aspects of the audio representation of the sound wave, commonly called the waveform, carry information about the size or material of the source, the type of excitation (e.g., knocking or rubbing) that generated it, or its perceived timbre? What is the acoustic basis of perceived dissimilarities, such as those between different instruments, different registers of the same instrument, and different players playing the same instrument? This chapter examines how differences in timbre manifest themselves in the audio signal and how such information can be extracted computationally from different signal representations in the form of *audio descriptors* to acoustically characterize timbre in music, speech, and environmental sounds.

Today timbre is understood from two perceptual viewpoints: as a sensory quality and as a contributor to source identity (Siedenburg and McAdams 2017). In the former, two sounds can be declared qualitatively dissimilar without bearing source-cause associations. In the latter, timbre is seen as the primary perceptual vehicle for the recognition and tracking over time of the identity of a sound source. Both approaches consider timbre as a very complex set of perceptual attributes that are not accounted for by pitch, loudness, duration, spatial position, and spatial characteristics such as room reverberation (Siedenburg, Saitis, and McAdams, Chap. 1). When timbre is viewed as qualia, its attributes underpin dissimilarity (McAdams, Chap. 2) and semantic ratings (Saitis and Weinzierl, Chap. 5). In the timbre-as-identity scenario, they facilitate sound source recognition (Agus, Suied, and Pressnitzer, Chap. 3). Further adding to its complex nature, timbre functions on different scales of detail (Siedenburg and McAdams 2017) in the sense that one sound-producing object can yield multiple distinct timbres (Barthet et al. 2010), and timbres from sound-producing objects of the same type but different "make" may differ substantially enough to affect quality judgements (Saitis et al. 2012). How informative is a given audio descriptor when examining different scales of timbre? What is the acoustic difference between a note played *pianissimo* and the same note played *fortissimo* or notes played in different registers on the same instrument?

Some of the most successful attempts to establish relationships between audio descriptors and perceptual aspects of timbre have resulted from multidimensional scaling (MDS) of pairwise dissimilarity ratings between musical instrument sounds (Grey and Gordon 1978; McAdams et al. 1995). Descriptors calculated from temporal and spectrotemporal representations of the audio signal are typically correlated with the dimensions of MDS *timbre spaces* to capture the acoustic cues underpinning the mental representation of timbre (McAdams, Chap. 2). Beyond psychoacoustics and music psychology, extracting quantitative descriptors potentially related to timbre from audio signals is an important part of the music information retrieval (MIR) discipline (Casey et al. 2008; Levy and Sandler 2009). The MIR task most relevant to timbre per se is musical instrument classification, which relies on an ensemble of descriptors associated with both the excitation-filter model and time-frequency representations to classify musical instrument sounds. However, the way audio descriptors are approached by MIR diverges from psychology due to differences in epistemic traditions and scientific goals between the two disciplines (Siedenburg et al. 2016a), a point discussed further in Sect. 11.4.1.

In MIR, descriptors are more commonly referred to as *features*. In psychology, features are discrete whereas dimensions are continuous (Peeters et al. 2011). In multimedia, features are perceptual by nature and descriptors are representations of features with specific instantiations (i.e., values) associated with data (Nack and Lindsay 1999). Pitch, for instance, is a feature of periodic sounds; fundamental frequency f_0 is a possible descriptor of pitch and $f_0 = 440$ Hz is the corresponding descriptor value. In MIR, features are extracted from the audio independently of the intrinsic nature of the information they represent (Casey et al. 2008). As such, a chord, the melody, and even the spectral envelope can be a feature. Following Peeters et al. (2011), the term *descriptor* is adopted here to disambiguate the concept of extracting information from the audio signal to describe its content.

Key questions that arise in working with audio descriptors of timbre include the following:

- What audio descriptors are appropriate for different tasks?
- What is the relation between the information captured by the descriptor and its usefulness?
- What is the relation between audio descriptors and perceptual, semantic, and cognitive aspects of timbre?
- What temporal information is important for timbre and how should it be represented with descriptors?
- How do we deal with timbral dimensions that covary with other perceptual dimensions?

Attempting to provide some answers to these questions, this chapter lays out a pathway into audio descriptor design and application. Section 11.2 presents basic audio representations that serve as a starting point for the extraction of audio descriptors presented in Sect. 11.3. Subsequently, Sect. 11.4 explores important applications of audio descriptors in the domain of timbre psychoacoustics, sound meta-description, musical orchestration, and sound morphing. Section 11.5 closes with a discussion of deep learning in automatic audio description and promising avenues for future research.

11.2 Representations of Audio Signals

This section introduces basic mathematical representations of audio from which audio descriptors can be extracted, which is not intended as a thorough explanation with a full mathematical treatment but rather as a general intuitive overview of the main concepts involved. The waveform (see Figs. 11.1–11.3) represents the pattern of pressure oscillations of a sound wave. Positive amplitude corresponds to compression and negative amplitude represents rarefaction. In digital audio, the discrete representation of sound waves is obtained by sampling continuous waveforms. Specifically, a discrete waveform is the result of sampling its analog counterpart at regular time intervals. The waveform contains all of the information carried by the sound wave it represents, but the waveform itself is seldom useful as a representation from which to extract perceptually meaningful information about timbral attributes or to categorize the sound source.

Figures 11.1–11.3 illustrate a typical sequence of steps taken to transform a waveform into a representation suitable for audio content description. The waveform is first windowed into time frames and the spectrum of each frame is obtained with the discrete Fourier transform (DFT). Descriptors are then computed globally or for each time frame. Details of the different steps are discussed below and in Sect. 11.3. To hear the sounds used in Figs. 11.1–11.3, go to the sound files "music.mp3", "speech.mp3", and "water.mp3".

11.2.1 Short-Time Fourier Transform and Spectrogram

The DFT is the standard method to obtain a representation of the frequency decomposition of the waveform called the frequency spectrum (Jaffe 1987a, b). The discrete frequencies of the DFT are linearly spaced, which means that adjacent frequency samples are separated by a constant interval called a *frequency bin*.

The short-time Fourier transform (STFT) analyzes a signal in terms of time *and* frequency by viewing it through successive overlapping windows, as depicted in Figs. 11.1–11.3, and then taking the DFT of each windowed frame (Portnoff 1980). The effect of the window is to concentrate the information in a short temporal frame.

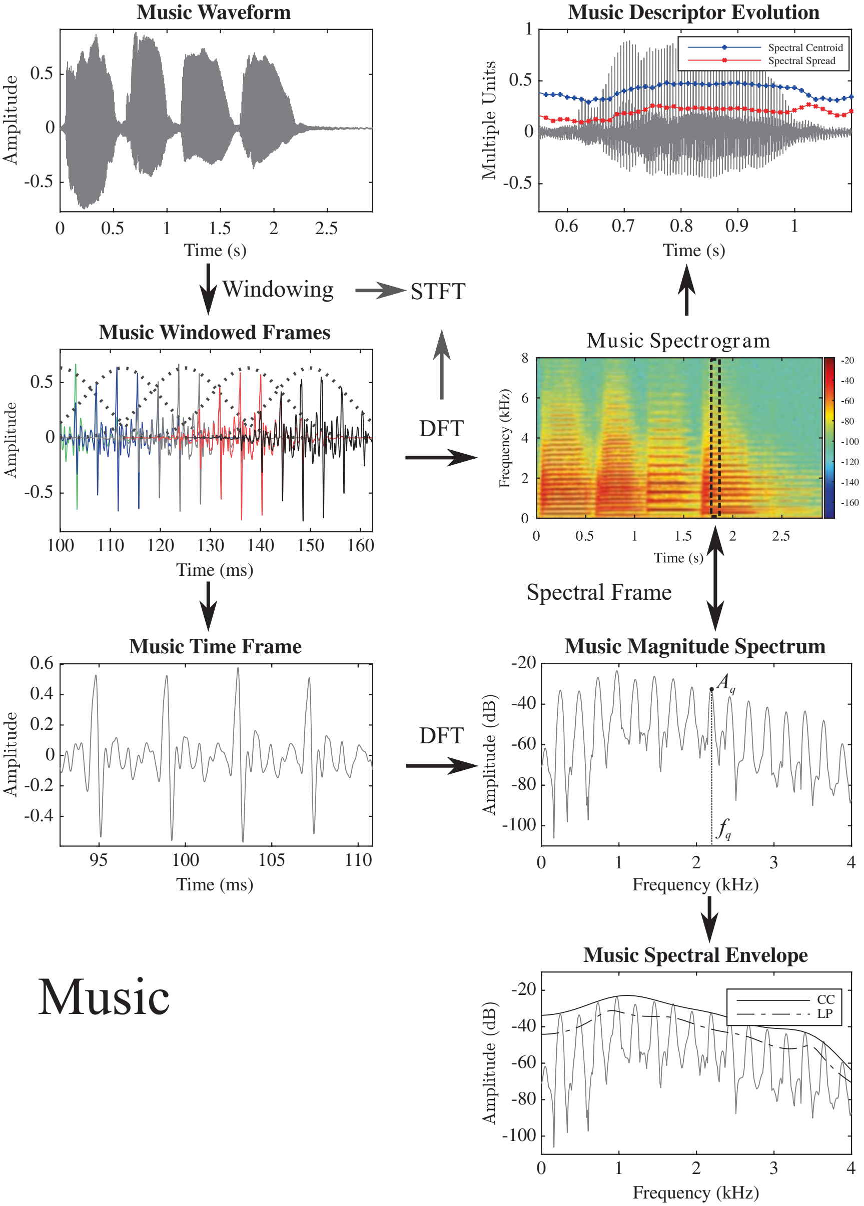

Fig. 11.1 Illustration of the sequence of steps to extract audio content descriptors from music. The *music* excerpt comprises four isolated notes (B3, G3, A4, and D4) of the monophonic trumpet track from a multitrack recording of Beethoven's Symphony No. 5. To hear the sounds, go to the sound file "music.mp3". The arrows indicate the connections between the panels. The extraction of descriptors from time-frequency representations is illustrated going *counter-clockwise* from the panel labeled *waveform* and the extraction of descriptors from the excitation-filter model is illustrated going *clockwise* from the panel labeled *time frame*. Abbreviations: A_q, amplitude of qth partial; *CC*, cepstral coefficients; *DFT*, discrete Fourier transform; f_q, frequency of qth partial; *LP*, linear prediction; *STFT*, short-time Fourier transform

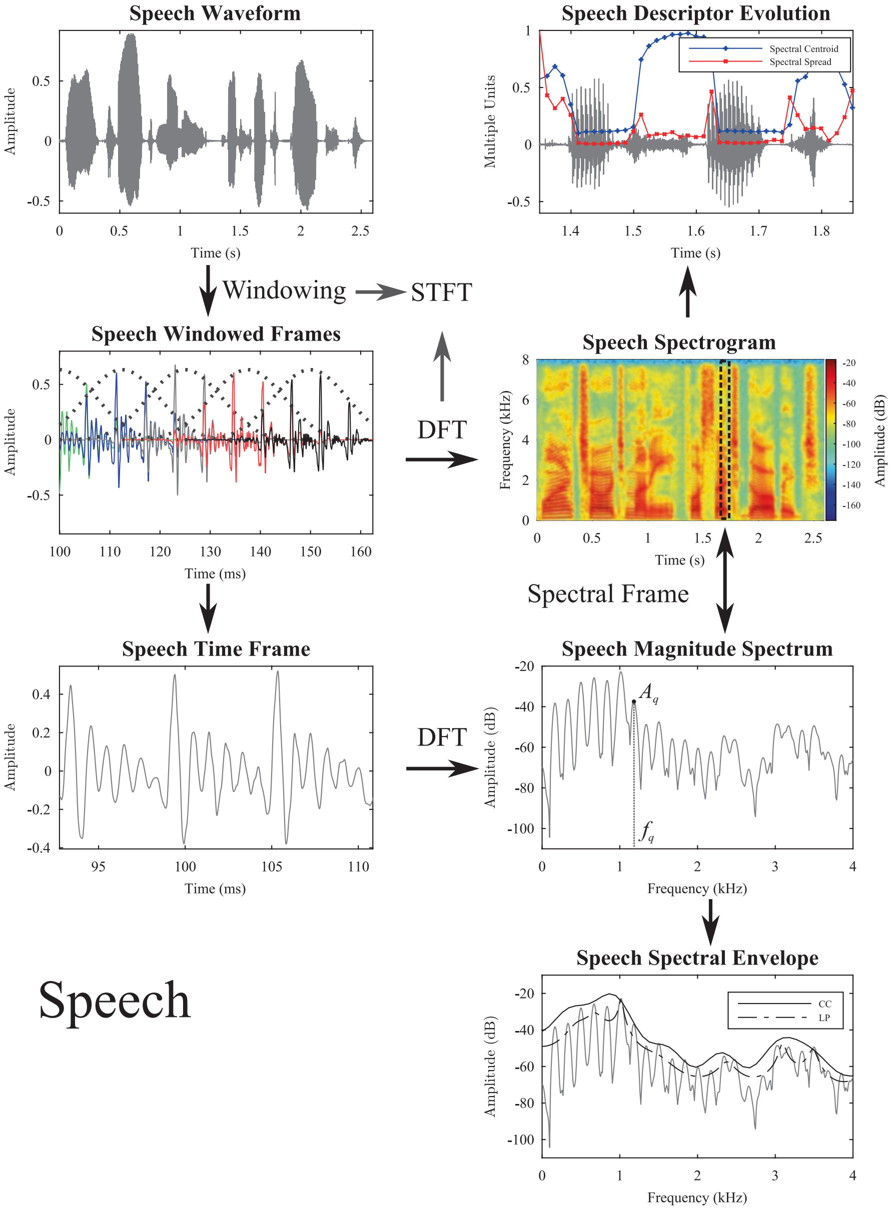

Fig. 11.2 Illustration of the sequence of steps to extract audio content descriptors from speech. The speech utterance consists of a 59-year-old white female speaker (U.S. southern accent) saying */why charge money for such garbage/*, sound file "speech.mp3". The extraction of descriptors from time-frequency representations is illustrated going *counter-clockwise* from the panel labeled *waveform* and the extraction of descriptors from the excitation-filter model is illustrated going *clockwise* from the panel labelled *time frame*. Abbreviations: A_q, amplitude of qth partial; *CC*, cepstral coefficients; *DFT*, discrete Fourier transform; f_q, frequency of qth partial; *LP*, linear prediction; *STFT*, short-time Fourier transform

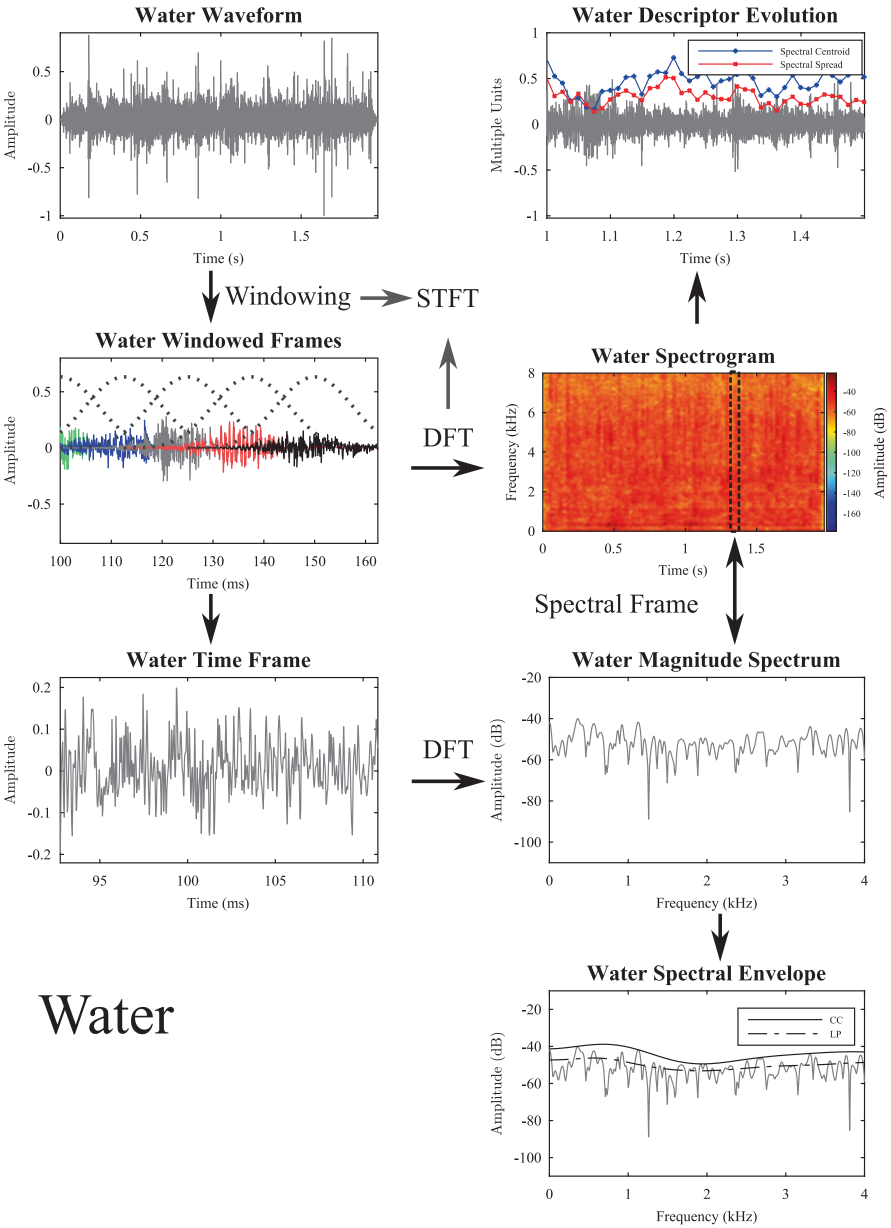

Fig. 11.3 Illustration of the sequence of steps to extract audio content descriptors from environmental sounds. The sample sound is running water, sound file “water.mp3”. The extraction of descriptors from time-frequency representations is illustrated going *counter-clockwise* from the panel labeled *waveform* and the extraction of descriptors from the excitation-filter model is illustrated going *clockwise* from the panel labelled *time frame*. Abbreviations: A_q, amplitude of qth partial; *CC*, cepstral coefficients; *DFT*, discrete Fourier transform; f_q, frequency of qth partial; *LP*, linear prediction; *STFT*, short-time Fourier transform

Figs. 11.1–11.3 show the *spectrogram*, a visualization of the STFT in which the magnitude spectrum of each DFT is plotted against time, while amplitude information (in dB) is mapped to color intensity. The STFT has become the de facto analysis tool for various speech and music processing tasks. However, the STFT has notorious limitations for spectral analysis, mainly due to the constant length of the window.

The STFT is inherently limited by the *Fourier uncertainty principle*, a mathematical relation stating that a function and its Fourier transform cannot both be sharply localized (Jaffe 1987b). For audio processing, this implies that there is a fundamental tradeoff between time and frequency information. The constant length of the window in the STFT results in fixed temporal and spectral resolutions. Intuitively, frequency is a measure of the number of periods (or cycles) per unit time. Longer windows span more periods, which increases the accuracy in frequency estimation while simultaneously decreasing the temporal localization of the measurement. Therefore, time-frequency uncertainty is at the core of Fourier analysis and only a priori knowledge about the analyzed signal type and about the spectral properties of the window used (Harris 1978) can help choose the most appropriate spectral analysis tools for a specific application.

11.2.2 Constant Q Transform

The STFT can be interpreted as a filter bank with constant bandwidth and linear separation of the center frequency of each filter (Portnoff 1980; Dolson 1986). The constant bandwidth of each filter is a direct consequence of the fixed window length, whereas the linear separation of their center frequencies is due to the constant frequency bins of the DFT. However, the frequency intervals of Western musical scales are geometrically spaced (Brown 1991), so the frequency bins of the STFT do not coincide with the musical notes of Western musical scales. Additionally, the constant bandwidth of the STFT imposes a tradeoff in time-frequency resolution, where a window length naturally results in better spectral resolution for higher frequencies at the cost of poorer temporal resolution. In practice, each octave would require a different window length to guarantee that two adjacent notes in the musical scale that are played simultaneously can be resolved. The *constant Q transform* exploits a nonlinear frequency separation with an adaptive window length to yield a more compact representation of Western musical scales (Brown 1991). The quality factor of a resonator, denoted Q, is defined as the resonance frequency divided by the bandwidth of the resonator. The resonance frequency is the frequency at which the peak gain occurs, whereas the bandwidth is the frequency range around the resonance frequency where the gain is above a predefined threshold. The higher the Q, the narrower and sharper the peak is.

The constant Q transform can be calculated similarly to the DFT with geometrically spaced frequency bins and frames with lengths that depend on the analysis frequency. For musical applications, the frequency separation can be based on the musical scale with the semitone spacing of the equal tempered scale. A constant Q in the frequency domain corresponds to a frame length that is inversely proportional to frequency because the constant Q transform is designed to span the same number of periods inside each time frame. Thus, the constant Q transform is equivalent to a filter bank with adaptive bandwidths and nonlinear center frequencies in which the center frequencies can be aligned with the musical scale and the bandwidths are proportional to the center frequencies to yield a similar spectral resolution across all octaves.

Despite being useful for the spectral analysis of Western music, the original constant Q transform algorithm (Brown 1991; Brown and Puckette 1992) remained less popular than the STFT for two main reasons. Firstly, the constant Q transform was computationally inefficient compared to the fast Fourier transform (FFT) commonly used to calculate the STFT. Secondly, the original constant Q transform (Brown and Puckette 1992) was not invertible—it allowed sound analysis but not resynthesis. Recently, however, an efficient real-time implementation of a fully invertible constant Q transform was made possible using the concept of Gabor frames (Holighaus et al. 2013).

11.2.3 Auditory Filter Banks

The concepts of auditory filter banks and critical bands of human hearing are closely related to spectrum analysis over nonlinear frequency scales. Auditory filter banks model the acoustic response of the human ear with a bank of nonuniform bandpass filters whose bandwidths increase as the center frequency increases (Lyon 2017). Critical bands correspond to equal distances along the basilar membrane and represent the frequency bands into which the acoustic signal is split by the cochlea. Zwicker (1961) proposed the *Bark scale* to estimate the value of the first 24 critical bands as a function of center frequency based on empirical measurements using two-tone masking of narrowband noise. The Bark scale is approximately linear for frequencies below about 500 Hz and close to logarithmic at higher frequencies. Later, Glasberg and Moore (1990) suggested the *equivalent rectangular bandwidth* (ERB) scale for critical band estimation based on measurements using notched-noise masking. The ERB of a given auditory filter is defined as the bandwidth of a rectangular filter with similar height (peak gain) and area (total power) as the critical band it models. The ERB values are similar to those obtained by the Bark scale for center frequencies above 500 Hz, but they are markedly narrower at lower frequencies and thus more consistent with critical bandwidths measured with the more precise notched-noise method.

Gammatone filters are a popular choice to model the shape and frequency response of auditory filters because of their well-defined impulse response. A gammatone is a simple linear filter defined in the time domain as a waveform with an amplitude envelope having the shape of a gamma distribution. Patterson et al. (1992) showed that certain gammatone shapes provide a nearly perfect approximation to the measured human auditory filter shapes. A more precise approximation is obtained by *gammachirp filters*, in which the sinusoid is replaced by a monotonically frequency-modulated signal (i.e., a chirp) (Irino and Patterson 1997). Compared to the STFT, ERB-spaced gammatone filter banks offer a physiologically more accurate representation of the audio signal from which to extract spectral descriptors (Peeters et al. 2011; Siedenburg et al. 2016b). Nevertheless, the use of auditory filter banks in acoustic analysis for timbre remains less widespread than the STFT or cepstrum-based techniques, which are more straightforward to implement and are perfectly invertible.

11.2.4 Sinusoidal Modeling

Sinusoidal models (McAulay and Quatieri 1986) are a convenient representation of sounds that feature periodicity, such as musical instrument sounds and speech (see Figs. 11.1, 11.2) under the assumption that the sinusoids capture locally periodic oscillations in the waveform. In essence, sinusoidal models represent spectral peaks with sinusoids because the DFT of a sinusoid appears as a peak in the magnitude spectrum (Jaffe 1987b). The *time frame* panels show that musical instruments (Fig. 11.1) and speech (Fig. 11.2) feature relatively stable periodic oscillations (locally), whereas environmental sounds rarely do (Fig. 11.3). The amplitude and frequency of each spectral peak (see the *magnitude spectrum* panels in Figs. 11.1–11.3) are estimated for each frame (McAulay and Quatieri 1986). The partials are called harmonics when their frequencies are integer multiples of a fundamental frequency. The sum of all time-varying amplitudes of the partials gives the temporal envelope of the sound (see Fig. 11.4).

11.2.5 Temporal Envelope

The temporal amplitude envelope follows fluctuations of the amplitude of a signal. Mathematically, it is possible to express a signal as a combination of a slowly varying envelope and a rapidly varying carrier signal. The temporal envelope and time-varying phase of this representation of the signal are useful in audio descriptor extraction because they model amplitude and phase modulations, respectively (Elhilali, Chap. 12). Tremolo and vibrato are also intrinsically related to these parameters. For example, Regnier and Peeters (2009) proposed to use vibrato to automatically detect a singing voice in polyphonic music.

The *Hilbert transform* and the closely related analytic signal are useful tools to estimate the temporal envelope without prior sinusoidal modeling (Peeters et al. 2011). A fundamental property of the DFT is behind the connection between the original signal and the analytic signal derived from it. The DFT of a real signal is complex and its magnitude spectrum is symmetric around the frequency axis, as

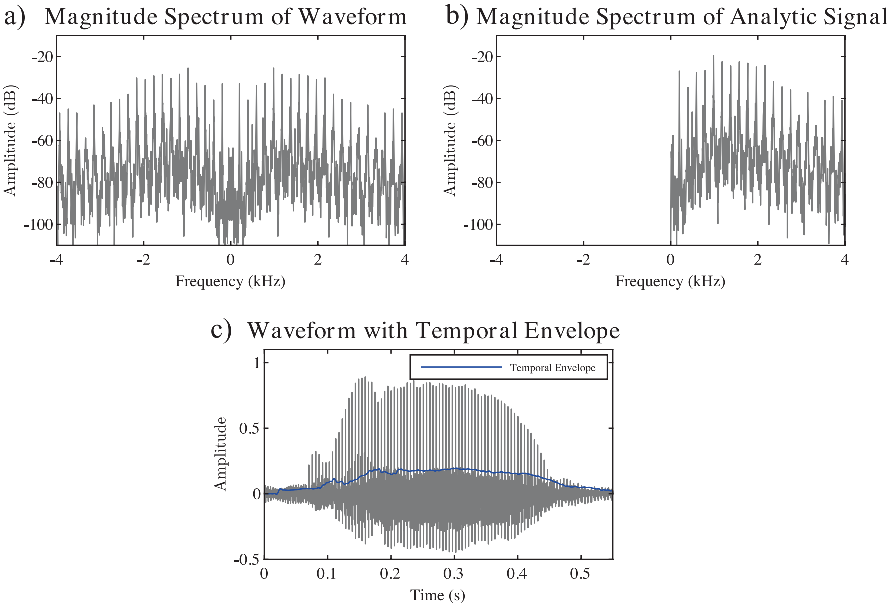

Fig. 11.4 Illustration of the analytic signal method to estimate the temporal amplitude envelope: (**a**) the magnitude spectrum of one of the trumpet notes from Fig. 11.1; (**b**) the magnitude spectrum of the analytic signal associated with (**a**); (**c**) the original waveform and the corresponding temporal envelope

shown in Fig. 11.4a. Mathematically, the property of symmetry means that the magnitude spectrum has a negative frequency component that has no physical interpretation. However, removing the negative frequencies and breaking the symmetry (see Fig. 11.4b) results in a spectrum that does not correspond to the original real signal anymore. In fact, the inverse DFT of the spectrum shown in Fig. 11.4b is a complex signal called the *analytic signal*, whose real part is the original signal and whose imaginary part is the Hilbert transform of the original signal. The temporal amplitude envelope can be calculated as the low-pass filtered magnitude of the analytic signal (Caetano et al. 2010; Peeters et al. 2011). Fig. 11.4c shows one of the trumpet notes seen in Fig. 11.1 with the temporal envelope calculated with the Hilbert transform. The Hilbert transform figures among the most widely used methods to estimate the temporal envelope, but it is hardly the only one (Caetano et al. 2010).

11.2.6 Excitation-Filter Model and Convolution

The excitation-filter model, also called the source-filter model (Slawson 1985; Handel 1995), offers a simple yet compelling account of sound production whereby a driving force, the excitation, causes a physical object, the filter, to vibrate. Here the term *excitation* is preferred over *source* to avoid potential confusion with the source of a sound, such as a musical instrument or a person. The physical properties of the vibrating object cause it to respond differently to different frequencies present in the excitation. Consequently, the vibrating object acts as a filter on the excitation, attenuating certain frequencies while emphasizing others. For example, a knock on a door is a short abrupt driving force that causes the door to vibrate. The sound from a wooden door is different from the sound from one with similar dimensions made of glass or metal due to their different material properties. Bowing the strings of a violin will cause its body to vibrate; the shape, size, and material of the violin body are responsible for the unique sonority of the instrument. Similarly, air through the lungs causes the vocal folds to vibrate, and the vocal tract shapes these vibrations into the unique timbre of a person's voice.

The interaction between the properties of the excitation and those of the vibrating object can be interpreted as filtering, which is mathematically expressed as a *convolution*. The Fourier transform is the key to understanding why convolution is mathematically equivalent to filtering because convolution in the time domain becomes multiplication in the frequency domain (Jaffe 1987b). This property of convolution is extremely useful for the analysis of sounds and the extraction of audio content descriptors of timbre in light of the excitation-filter model. In particular, the filter component (or transfer function) models how the physical properties of the vibrating object respond to the excitation in the frequency domain. The contributions of the excitation and filter can theoretically be isolated in the frequency domain and inverted, bringing the frequency spectrum back to the time domain. In the time domain, the transfer function is called the *impulse response* and is essentially a model of the physical properties of the vibrating object. Consequently, the impulse response carries information intrinsically related to timbre perception that can be used to extract audio descriptors of timbre. Section 11.3.3 explores some of the most widely used audio descriptors of timbre based on the excitation-filter model.

11.3 Extraction of Timbre Descriptors

The raw information provided by audio signal representations such as the STFT and the excitation-filter model is usually not specific enough to describe salient aspects of timbre. Therefore, a plethora of techniques for extracting timbre-relevant descriptors from these representations has been proposed in the field of audio content analysis. Some audio descriptors are extracted from generic time-frequency representations and are later found to capture aspects of timbre perception, whereas others are based on the excitation-filter model and commonly describe physical properties of the sound source. In general, audio descriptors can represent global or local aspects of sounds. *Global descriptors* only have one value for the entire duration of a sound, whereas *local descriptors* are commonly calculated for every frame (see Figs. 11.1–11.3) and result in a time series.

Additionally, descriptors can be categorized as temporal, spectral, or spectrotemporal (Peeters et al. 2011). *Temporal descriptors* exclusively capture temporal aspects of sounds and are generally global. Some are computed directly from the waveform, but most are typically extracted from the temporal energy envelope (Sect. 11.2.5). *Spectral descriptors* capture local features of the frequency content regardless of the surrounding frames. Spectral descriptors also have an alternative harmonic version calculated from the sinusoidal model. Finally, *spectrotemporal descriptors* capture spectral changes relative to the previous or next frames. Thus, spectrotemporal descriptors attempt to incorporate time as relative local spectral changes throughout the duration of a sound.

Section 11.3.1 addresses temporal descriptors and Sect. 11.3.2 covers descriptors extracted from time-frequency representations. Section 11.3.3 focuses on descriptors based on the excitation-filter model, Sect. 11.3.4 explores the temporal dynamics of the time series of descriptors, and Sect. 11.3.5 discusses information redundancy among descriptors.

11.3.1 Temporal Descriptors

The zero-crossing rate is a measure of how many times the waveform changes sign (i.e., crosses the zero axis). In general, periodic sounds have a smaller zero-crossing rate than noisier sounds, so the zero-crossing rate can be used in voice activity detection, voiced-unvoiced decisions for speech, and even in the classification of percussive sounds (Peeters et al. 2011), although there is no straightforward perceptual interpretation of the zero-crossing rate (Siedenburg et al. 2016a).

The temporal envelope is used to extract temporal descriptors such as tremolo (Peeters et al. 2011), the temporal centroid, and attack time. The *temporal centroid* is the temporal counterpart of the spectral centroid (see Sect. 11.4.2). Percussive sounds have a lower temporal centroid than sustained sounds. McAdams et al. (2017) found that a lower (i.e., earlier) temporal centroid correlated strongly with the valence of musical affect carried by the timbre of musical instrument sounds.

The *attack time* is the time between the onset of a sound and its more stable part. In musical instruments, for example, the attack time accounts for the time the partials take to stabilize into nearly periodic oscillations. Percussive musical instruments, such as the xylophone, feature short attack times with sharp onsets, whereas sustained instruments, such as the tuba, feature longer attack times. The attack time of a waveform can be estimated with models such as the weakest effort method (Peeters et al. 2011) or the amplitude/centroid trajectory model (Hajda 2007; Caetano et al. 2010). The *weakest effort method* uses signal-adaptive energy thresholds instead of fixed energy levels to estimate the beginning and end of the attack from the temporal envelope. The *amplitude/centroid trajectory model* uses spectrotemporal information from both the temporal envelope and the temporal evolution of the spectral centroid to segment musical instrument sounds. Section 11.3.3 will delve deeper into the amplitude/centroid trajectory model and the timbre descriptors used therein. The attack time consistently arises as one of the most salient dimensions in timbre spaces from MDS studies (Grey 1977; Siedenburg et al. 2016a).

McAdams et al. (1995) found the logarithm of the attack time among the most salient dimensions of perceptual dissimilarity.

Other common temporal descriptors (Peeters et al. 2011) include the slopes of the energy envelope during the attack and decrease segments, the effective duration, and the temporal modulation of energy over time (i.e., tremolo). Energy modulation is calculated either from the temporal evolution of the amplitudes of isolated partials across frames or from the temporal envelope.

11.3.2 Time-Frequency Representations and Audio Descriptors

11.3.2.1 Spectral Descriptors

Spectral descriptors are typically calculated for each frame of a time-frequency representation such as the STFT (see Figs. 11.1–11.3). Descriptors of spectral shape characterize the overall spectral distribution of sounds and are calculated as if the STFT magnitude spectrum were a probability distribution. Peeters et al. (2011) remarked that spectral descriptors can use different spectral scales such as magnitude, power, or log.

The spectral shape descriptors, calculated similarly to the standardized moments of the frequency spectrum, are the spectral centroid, spectral spread, spectral skewness, and spectral kurtosis. The *spectral centroid* is the amplitude-weighted mean frequency. It is measured in Hertz (Hz) and is analogous to the center of mass, so it can be interpreted as the center of balance of spectral energy distribution or the frequency that divides the spectrum into two regions with equal energy. The spectral centroid often appears among the most salient dimensions of timbre spaces (see McAdams, Chap. 2), and it is interpreted as capturing the "brightness" of a sound (Grey and Gordon 1978; McAdams et al. 1995). Sounds described as bright, such as a brassy trombone note, have higher spectral centroids because they feature more spectral energy in high frequency regions (see Sect. 11.4.1). *The spectral spread* measures the spread of spectral energy around the spectral centroid. It is related to the bandwidth of a filter, so a brighter sound will have a higher spectral spread than a duller sound. *Spectral skewness* is a measure of asymmetry of spectral energy around the spectral centroid. Negative values indicate more spectral energy concentrated at frequencies lower than the spectral centroid, positive values indicate more energy at higher frequencies than the centroid, and zero indicates energy symmetry around the centroid. Finally, *spectral kurtosis* is a measure of the flatness of the spectral distribution of energy compared to a normal distribution. A negative value of spectral kurtosis indicates a distribution of spectral energy flatter than the normal distribution, whereas a positive value indicates the opposite.

Spectral flux or *spectral variation* (Casey et al. 2008; Peeters et al. 2011) is considered a spectrotemporal descriptor because it captures local spectral change over time. Essentially, it measures the spectral difference of the current frame relative to the previous frame. Compared to attack time and spectral centroid, the correlation of spectral flux with listener ratings of timbre similarity has been less consistent. While in some studies the third dimension of the MDS timbre space does correlate moderately well with the time-varying spectral flux (McAdams et al. 1995), in others it correlates better with static descriptors of *spectral deviation* (deviation of partial amplitudes from a global, smoothed spectral envelope; Krimphoff et al. 1994) or *spectral irregularity* (attenuation of even harmonics; Caclin et al. 2005).

Several other spectral descriptors appear in the literature (Peeters et al. 2011), many of which capture similar properties of the spectrum. However, there is little consensus about their usefulness or even their relationship with timbre (see Sect. 11.4.1).

11.3.2.2 Harmonic Content

Most spectral descriptors also have a harmonic version calculated by simply replacing the spectral magnitude with the amplitudes of the sinusoidal model (see the *magnitude spectrum* panels in Figs. 11.1–11.3), such as the harmonic energy (Peeters et al. 2011). However, some descriptors capture information specifically related to the oscillatory modes of the signal, commonly called *partials*. Figs. 11.1–11.3 highlight the differences in both time and frequency domains for sounds of musical instruments, speech, and environmental sounds (represented by running water in Fig. 11.3). The *time frame* panels reveal that both musical instruments and speech feature relatively stable oscillations in some regions (except where changes, such as note transitions, are happening), whereas the running water sound is noisy. Oscillations in the time domain appear as spectral peaks in the frequency domain. The magnitude spectrum of the musical instrument shows prominent spectral peaks across the entire frequency range of 0–4 kHz. For speech, the spectral peaks are less prominent beyond approximately 2.2 kHz. Finally, the magnitude spectrum for the water sound shows a relatively flat distribution of spectral energy typical of noise.

A fundamental result from Fourier analysis (Jaffe 1987a) reveals that the spectrum of a perfectly periodic waveform is perfectly harmonic. However, neither speech nor musical instrument sounds are perfectly periodic. Consequently, neither type has a spectrum that features perfectly harmonic spectral peaks. This can be quantified with the descriptor *inharmonicity*, based on the sinusoidal model (see Sect. 11.2.4). Inharmonicity measures the deviation of the frequencies of the partials from pure harmonics, calculated as the normalized sum of the differences weighted by the amplitudes (Peeters et al. 2011). Sustained musical instruments, such as those from the woodwind (e.g., flute, clarinet, bassoon, and oboe), brass (e.g., trumpet, trombone, and tuba), and string (e.g., violin, viola, and cello) families, produce sounds whose spectra are nearly harmonic (Fletcher 1999). Percussion instruments (e.g., cymbals and timpani) are considered inharmonic, whereas others (e.g., bells or the piano) feature different degrees of inharmonicity (Fletcher 1999; Rigaud & David 2013). The spectrum of the piano, for example, has partials whose inharmonicity is proportional to the partial number. So, the higher the frequency, the greater is the deviation from the harmonic series (Rigaud & David 2013). This characteristic inharmonicity is an essential property of the timbre of the piano. Features that are specific to certain musical instruments, commonly called *specificities*, are directly related to timbre perceptions of these instruments. For instance, the timbre of clarinet sounds, described as "hollow" (McAdams et al. 1995), can be linked to spectral energy predominantly concentrated around odd harmonics. The *odd-to-even harmonic energy ratio* (Peeters et al. 2011) is a descriptor that quantifies this particular specificity.

Pollard and Jansson (1982) proposed a three-dimensional representation of timbre dubbed *tristimulus*. Each dimension of the tristimulus representation contains the loudness of a group of partials (i.e., how much energy each group contributes to the overall spectrum). The first dimension has the fundamental frequency, the second dimension includes partials two to four, and the third dimension contains the rest of the partials from the fifth to the highest. Pollard and Jansson (1982) used the tristimulus method to represent the temporal evolution of musical instrument sounds and revealed variations in timbre with time, especially between the attack transients and the steady state with its more stable oscillatory behavior. Section 11.3.4 will explore further the temporal evolution of descriptors and timbre.

11.3.3 The Excitation-Filter Model and Audio Descriptors

There are several descriptors of timbre based on the excitation-filter model of sound production (introduced in Sect. 11.2.6). These descriptors typically capture information related to the filter component of the model, which is responsible for the relative distribution of spectral energy. Perceptually, the relative energy of spectral components is directly related to timbre and is sometimes called sound color (Slawson 1985). When associated with the excitation-filter model, the spectral envelope (see the *spectral envelope* panels in Figs. 11.1–11.3) is commonly used to represent the filter component.

Descriptors of timbre based on the excitation-filter model commonly use the magnitude spectrum and discard the phase, autocorrelation coefficients being the quintessential example. Autocorrelation is a measure of self-similarity, whereby a signal is compared with its own past and future values. The autocorrelation and convolution operations share similarities that become more evident with the DFT (Jaffe 1987b). The autocorrelation coefficients are the time domain representation of the power spectral density (Jaffe 1987b; Brown et al. 2001), so they are related to the filter component. The relationship between autocorrelation coefficients and power spectral density is exploited further by linear prediction (Makhoul 1975).

11.3.3.1 Linear Prediction Coefficients

Linear prediction (Makhoul 1975) assumes that a signal can be described as a weighted linear combination of past values plus an external influence. The external influence accounts for the force exciting the vibrating object that generated the signal, whereas the vibrating object itself is not explicitly modeled. When the external influence is unknown, the signal can only be approximated by its past values. The model parameters can be estimated by minimizing the mean squared error. The solution yields the set of *linear prediction coefficients* (LPC) that best predict the next value of the signal given a specific number of preceding values in the least squared error sense, which is mathematically equivalent to using the autocorrelations to estimate the LPC (Makhoul 1975).

The LPC are commonly represented in the frequency domain with the Z-transform, which encodes essentially the same information as the Fourier transform (Jaffe 1987b) but in a more general framework. Similarly to the DFT, the Z-transform can also be interpreted as the frequency domain representation of the signal. The Z-transform of the linear prediction model explicitly reveals the frequency response of the vibrating object under the force that resulted in the sound spectrum. This frequency response is the filter component, commonly called the transfer function, and it fully characterizes the model of sound production under certain assumptions (Makhoul 1975).

Immediate physical and physiological interpretations for musical instrument sounds and speech can be derived from the LPC. For example, the LPC can be interpreted as a model of the resonances of the vocal tract in speech production (Makhoul 1975) because they encode the poles of the filter that approximates the original power spectral density. Linear prediction is commonly used to approximate the power spectrum with the spectral envelope (see the LP curve in the *spectral envelope* panels in Figs. 11.1–11.3), defined as a smooth curve that approximately connects the spectral peaks (Burred et al. 2010).

11.3.3.2 The Cepstrum

The cepstrum (Bogert et al. 1963; Childers et al. 1977) is intimately connected with the excitation-filter model because it was originally developed as a deconvolution method (Bogert et al. 1963). The excitation-filter model postulates that a waveform can be described mathematically as the convolution between the filter and the excitation. Deconvolution allows recovery of either the filter or the excitation from the waveform. In the frequency domain, convolution becomes multiplication (see Sect. 11.2.6) and deconvolution becomes inverting the result of the multiplication. Division is the simplest method when the DFT of either the excitation or the filter is available, allowing recovery of the other. However, in most practical applications, only the waveform resulting from the convolution between the excitation and the filter is available. In this case, the logarithm can be used to transform the multiplication operation into addition. If the terms of the resulting addition do not overlap in frequency, it is possible to isolate either one from the other by filtering. The *cepstrum* is the formalization of this deconvolution operation (Childers et al. 1977), which has found several applications in audio research, such as fundamental frequency estimation (Childers et al. 1977), spectral envelope estimation (Burred et al. 2010), wavelet recovery (Bogert et al. 1963), and musical instrument classification (Brown 1999; Herrera-Boyer et al. 2003). The *spectral envelope* panels in Figs. 11.1–11.3 show its estimation with the cepstrum (identified as *CC*).

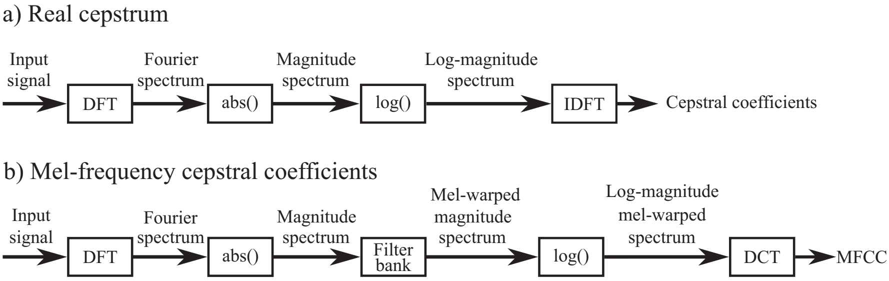

Fig. 11.5 Illustration of the sequence of steps to calculate cepstral coefficients with the real cepstrum (**a**) and mel-frequency cepstral coefficients (MFCC) (**b**) from the waveform. Abbreviations: *abs()*, absolute value; *DCT*, discrete cosine transform; *DFT*, discrete Fourier transform; *IDFT*, inverse discrete Fourier transform; *log()*, logarithm function; *MFCC*, mel-frequency cepstral coefficients

The real cepstrum can be defined as the inverse DFT of the logarithm of the magnitude of the DFT of the waveform. Fig. 11.5 illustrates the steps to obtain cepstral coefficients from a waveform (labeled *input signal*). The cepstral coefficients contain frequency information about the log magnitude spectrum similarly to how the LPC encode the resonances of a transfer function. In practice, these coefficients encode information about periodicity of the log magnitude spectrum at increasing cepstral frequencies, which were originally called “quefrencies” (Bogert et al. 1963), because they carry frequency information in time domain units. This unfamiliar symmetry was reflected in language by rearranging syllables of familiar terms from Fourier analysis. Particularly, “cepstrum” derives from spectrum and is pronounced *kepstrum*.

11.3.3.3 Mel-Frequency Cepstral Coefficients

Conceptually, the cepstral coefficients are closely related to the filter component of the excitation-filter model and of the ubiquitous mel-frequency cepstral coefficients (MFCC; mel is short for melody). Davies and Mermelstein (1980) introduced MFCC in the context of speech research. The MFCC can be viewed as a perceptually inspired variation of cepstral coefficients calculated as illustrated in Fig. 11.5. The MFCC filter bank uses triangular filters centered at frequencies given by the mel scale with a bandwidth proportional to the center frequency.

The perception of pitch allows listeners to order sounds on a scale from low to high along the same psychological dimension of melody (Hartmann 1996). A sound has a certain pitch if it can be reliably matched by adjusting the frequency of a sine wave of arbitrary amplitude (Hartmann 1996). The mel scale was derived by asking listeners to set the frequency of a test sine wave to obtain a pitch that was a fraction of the pitch of a reference sine wave across the entire audible frequency range (approximately between 20 Hz and 20 kHz). It is linear up to 1 kHz and logarithmic above 1 kHz. Stevens et al. (1937) concluded that the mel scale captures the concept of pitch height (i.e., higher or lower pitches) as opposed to pitch chroma (i.e., the octave-independent musical notes). The MFCC use the discrete cosine transform (commonly used in MPEG audio and image compression) instead of the DFT or the Z-transform commonly used for cepstral coefficients. Thus, MFCC are considered a particularly compact representation of the filter due to the compression properties of the discrete cosine transform, which results in most of the spectral shape being captured typically by the first thirteen coefficients.

The MFCC are ubiquitous not only in speech for tasks such as speaker recognition (On et al. 2006; Martínez et al. 2012) but also in MIR tasks such as musical instrument classification (Deng et al. 2008). Some results suggest that MFCC can also explain timbre. For example, Terasawa et al. (2005) compared MFCC, LPC, and tristimulus (see Sect. 11.3.2.2) representations to explain the pairwise perceptual dissimilarity ratings of sounds created with frequency-modulation synthesis. They found that the Euclidean distances between MFCC accounted for 66% of the variance and concluded that thirteen MFCC can be used as a model of timbre spaces. Horner et al. (2011) compared different error metrics to predict the discrimination performance of listeners for sounds synthesized with fixed fundamental frequency and variable spectral envelope. They found that the first twelve MFCC were sufficient to account for around 85% of the variance of data from human listeners.

11.3.4 Temporal Dynamics of Audio Descriptors

Many descriptors are calculated for every frame of time-frequency representations, such as the STFT, giving rise to a time series of descriptor values that characterizes the temporal evolution of each descriptor. The *descriptor evolution* panels in Figs. 11.1–11.3 show the temporal evolution of the spectral centroid and spectral spread, revealing local variations corresponding to changes such as note transitions. However, most applications, such as musical instrument classification, require one single value of each descriptor that would be representative of the entire sound duration. Commonly, the time average of each descriptor is used for each sound, resulting in a descriptor vector. Descriptors such as the spectral centroid are unidimensional, whereas others, such as MFCC, are multidimensional. Therefore, descriptor vectors discard all information about the temporal variation of descriptors.

The simplest way to include more information than the time average of the descriptors is to use a set of summary statistics such as mean, standard deviation (or variance), minimum, and maximum values (Casey et al. 2008). Peeters et al. (2011) found that robust summary statistics had a greater impact than the audio representation on the descriptors. Specifically, the median and the interquartile range captured distinct aspects of the signals. McDermott et al. (2013) suggested that environmental sounds are recognized by summary statistics alone because the temporal information in environmental sounds can be captured by summary statistics. However, the temporal structure inherent to speech and musical sounds requires encoding temporal information in different ways.

The first and second derivatives with respect to time (of the time series of descriptor values) are another popular approach to include temporal information in the descriptor vector. It is particularly common to use MFCC and their first and second temporal derivatives, called delta and delta-delta coefficients, respectively (De Poli and Prandoni 1997; Peeters et al. 2011). However, the delta and delta-delta coefficients are usually added to the descriptor vector as extra dimensions assumed to be independent from the descriptor values. Consequently, the information contained in the time series of descriptor values is not fully exploited. For example, Fig. 11.2 reveals that the spectral centroid of speech varies considerably between periodic and noisier segments. Similarly, for musical instruments, the temporal variation of descriptors follows musical events such as note transitions. The amplitude/centroid trajectory model (Hajda 2007) shown in Fig. 11.6 proposes to use the root-mean-squared amplitude envelope in conjunction with the temporal evolution of the spectral centroid to segment sustained sounds from musical instruments into attack, transition (so-called decay), sustain, and release portions. Fig. 11.6 shows the amplitude-centroid trajectory model used to segment notes from sustained musical instruments (Caetano et al. 2010). Segmentation of musical instrument sounds with the amplitude-centroid trajectory model yields better results for sustained instruments than percussive ones because sustained instruments fit the model better.

The use of a descriptor vector with the time average of each descriptor in each dimension is called the *bag of frames approach* because it treats the time series of descriptors as a global distribution, neglecting both the temporal variation and the sequential order of descriptor values (Levy and Sandler 2009; Huq et al. 2010). This approach can be successfully used to classify environmental sounds (Aucouturier

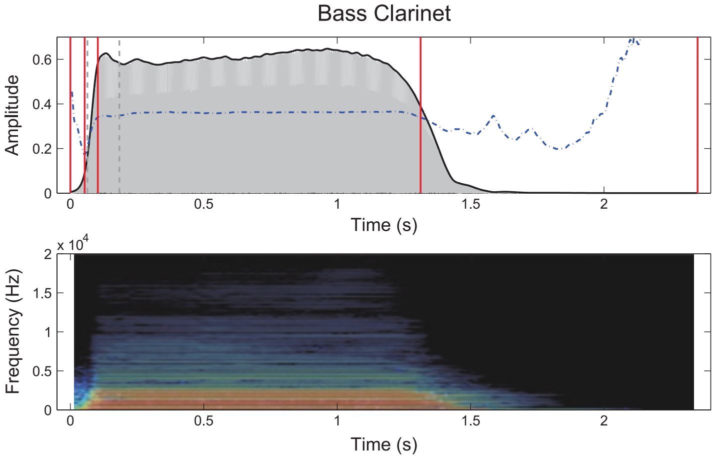

Fig. 11.6 Temporal segmentation of musical instrument sound with the Amplitude/Centroid Trajectory (ACT) model. Top panel: the full-wave rectified waveform outlined by the temporal amplitude envelope (*solid line*) and the temporal evolution of the spectral centroid (*dashed line*). The vertical bars mark the segments estimated with the ACT method. See text for an explanation of the segments. Bottom panel: the spectrogram of the waveform on the top. (Reproduced from Caetano et al. 2010; used with permission)

et al. 2007) with Gaussian mixture models representing the global distribution of MFCC. However, it is inappropriate for polyphonic music (Aucouturier et al. 2007) in which the sequence of events contains important information. In music, there is a clear hierarchical structure where higher levels of abstraction emerge from lower levels. For example, patterns of notes are organized into phrases, and rhythmic structure emerges from relative note durations. Aucouturier et al. (2007) speculated that the hierarchical structure of polyphonic music carries information on a more symbolic level than is captured by descriptors such as MFCC, requiring incorporation of information such as harmony and melody.

Temporal modeling of descriptors has been successfully applied in instrument classification and detection. Models of musical instrument sounds that rely on spectrotemporal representations are capable of capturing the dynamic behavior of the spectral envelope (Burred and Röbel 2010; Burred et al. 2010). Principal component analysis reduces the dimensionality of the model by projecting the time-varying parameters of the spectral envelopes onto a lower-dimensional space, such as the three-dimensional space shown in Fig. 11.7. The resultant prototypical temporal evolution of the spectral envelopes was modeled as a nonstationary Gaussian process and was shown to outperform MFCC for the classification of isolated musical instruments and to allow for instrument recognition in polyphonic timbral mixtures.

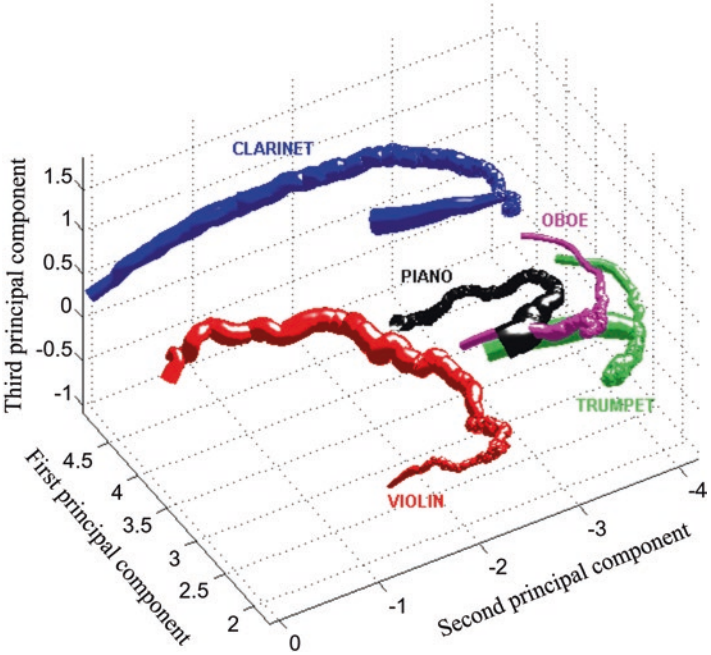

Fig. 11.7 Temporal evolution of the spectral envelope of musical instrument sounds. The temporal trajectory of the spectral envelope of the musical instruments indicated (*clarinet*, *oboe*, *piano*, *trumpet*, and *violin*) is shown in a three-dimensional representation obtained with principal component analysis. (Reproduced from Burred and Röbel 2010; used with permission)

11.3.5 Information Redundancy of Audio Descriptors

Descriptor vectors stack several descriptors under the assumption that each dimension is statistically independent from the others. While this assumption might hold true for some descriptors, such as MFCC, which are decorrelated by construction due to the discrete cosine transform (see Sect. 11.3.2.2), other descriptors are highly correlated. Peeters et al. (2011) investigated the correlation structure among descriptors extracted with alternative representations based on an analysis of over 6000 musical instrument sounds with different pitches, dynamics, articulations, and playing techniques. The authors observed that a change in the audio representation (e.g., STFT versus ERB-spaced filterbank versus harmonic content) had relatively little effect on the interdescriptor correlation compared to the change in the summary statistic computed on the time-varying descriptors, although no prediction of perceptual data was undertaken in that paper.

Several strategies have been proposed to decrease information redundancy in descriptor vectors. Among these, the most common ones fall generally into *descriptor selection* or *descriptor decomposition* strategies. Descriptor selection involves finding the subset of descriptors that is useful to build a good predictor (Huq et al. 2010) by eliminating descriptors that are either irrelevant or redundant. On the other hand, descriptor decomposition techniques apply transformations on the original space of descriptors that aim to maximize the information that is relevant for a task in the reduced space, such as the variance of the descriptors or the discriminability of classes. These transformations commonly involve projection or compression techniques, such as principal component analysis for the former and the discrete cosine transform for the latter. Descriptor decomposition techniques commonly distort the original representation in ways that can render interpretation more difficult. For example, principal component analysis results in linear combinations of the original dimensions that, in practice, render a perceptual interpretation of the results that is more arduous because each principal component accounts for more than one descriptor. Descriptor selection preserves the original meaning of the variables by preserving their original representation, ultimately offering the advantage of interpretability. At the same time, descriptor selection can lead to choices that seem arbitrary in that the selected descriptors may vary a great deal from one study to another.

11.4 Applications of Timbre Descriptors

Audio content descriptors find several applications that involve timbre description. Examples about the study of timbre psychoacoustics are discussed in Sect. 11.4.1; the multimedia content description interface also known as MPEG-7 is discussed in Sect. 11.4.2; computer-aided orchestration is discussed in Sect. 11.4.3; and musical instrument sound morphing guided by descriptors of timbre is discussed in Sect. 11.4.4.

11.4.1 Timbre Psychoacoustics

Audio signal descriptors have been central to the psychoacoustics of timbre, which seeks an explanation of timbre perception on acoustic grounds. Most of this research has used musical instrument sounds. A notable exception is the work by Zwicker and Fastl (1990), who presented acoustic models of sharpness, fluctuation strength, and roughness, which have been used mainly to characterize the sound quality of product sounds (see Lemaitre and Susini, Chap. 9). In the following discussion, three examples of using audio descriptors for psychoacoustics research will be addressed. These examples highlight the search for acoustic correlates of timbral brightness judgements and sound source recognition.

For musical sounds, two methods to study brightness perception can be distinguished. First, timbre space dimensions obtained via MDS of general dissimilarity judgements have consistently been interpreted as associated with the brightness of sounds (see McAdams, Chap. 2). Second, several studies have directly asked participants to rate the brightness of sounds and have correlated the resulting ratings with descriptor values. For instance, Schubert and Wolfe (2006) considered whether direct brightness ratings are better predicted by the absolute spectral centroid or the (supposedly pitch invariant) centroid rank (the centroid divided by the fundamental frequency). The latter predictor, however, failed to correlate significantly with subjective brightness, whereas the absolute centroid did.

Marozeau and de Cheveigné (2007) proposed a refined spectral centroid descriptor to model the brightness dimension of dissimilarity ratings. The approach was conceptually related to the sharpness descriptor by Zwicker and Fastl (1990) in that it relied on the computation of partial loudness in spectral bands (but the Zwicker model only insufficiently predicted brightness scaling data in Almeida et al. 2017). Specifically, the descriptor by Marozeau and de Cheveigné (2007) was obtained from partial loudness values calculated in ERB-spaced spectral bands obtained from gammatone filtering (see Sect. 11.2.3). An instantaneous spectral centroid was obtained through the integration across bands and the resulting time series was integrated over time by weighting with an estimate of instantaneous loudness (the sum over channels of partial loudness). In comparison to the linear spectral centroid descriptor, the refined brightness descriptor by Marozeau and de Cheveigné (2007) improved the amount of the explained variance with the perceptual data by 10% points up to 93%. Further analysis showed that it was much less affected by pitch variation compared to the more simplistic linear spectral centroid.

Fewer studies have used signal descriptors to address the acoustic features underlying sound source recognition and classification (see Agus, Suied, and Pressnitzer, Chap. 3). Ogg et al. (2017) modeled participant responses in a go/no-go categorization task of short sound excerpts varying in duration (12.5–200 ms). Three sound categories were tested: speech, musical instruments, and human environmental sounds generated by everyday objects (e.g., keys jingling), by objects of various materials impacting one another or being deformed (e.g., crumpling newspaper), and sounds generated by movements of liquid (fingers splashing) or rolling objects (marbles rolling down wood). Using exploratory regression analysis with timbre descriptors from the Timbre Toolbox (Peeters et al. 2011), the study characterized the acoustic features that listeners were using to correctly classify sounds. Globally, regression analysis for sounds from every target category indicated that listeners relied on cues derived from spectral, temporal, pitch, and "noisiness" information. Different sound categories required different sets of descriptors and weightings of regression coefficients. For instance, as the median spectral centroid value increased, listeners were more likely to categorize the stimuli as human environmental sounds and less likely to consider the sounds as coming from musical instruments. The descriptors "noisiness" and "spectral flatness" were associated with environmental and instrument responses, respectively.

Approaches such as these provide valuable starting points to reveal the most important acoustic features for a given psychophysical task from the plethora of available audio content descriptors. Giordano et al. (2012) further showed that audio descriptors can be applied to neuroimaging research (for neurophysiological details, see Alluri and Kadiri, Chap. 6). Following the approach of representational similarity analyses, they used descriptors to decode fMRI data recorded while participants listened to environmental sounds. They extracted descriptors based on pitch, loudness, spectral centroid, and harmonicity, and they computed dissimilarity matrices that contained the pairwise dissimilarity of stimuli according to these descriptors. Dissimilarity matrices were also derived from the imaging data, specifically, from the response of each voxel in a region of interest. Then, correlation of the neurophysiological and the acoustic dissimilarity matrices resulted in maps that indicated the association of the activity in a given voxel to a specific acoustic property. Hence, this approach can infer the brain areas associated with the processing of low-level acoustic properties represented by the audio descriptors.

These examples indicate that a variety of psychophysical and even psychophysiological questions on timbre can benefit from a deeper involvement with audio descriptors, which can be easily computed today (Peeters et al. 2011). At the same time, the correlational nature of the approach warrants rigorous confirmatory studies to circumvent the strong mutual covariance of descriptors.

More generally, it seems important to acknowledge that work on timbre-related audio content descriptors is at the crossroads of distinct academic fields, including MIR, music cognition, and psychoacoustics. Hence, it is important to appreciate the distinct epistemic traditions and objectives that are ingrained in these fields (Siedenburg et al. 2016a). Music information retrieval is a task-oriented discipline rooted in applied computer science and machine learning and, therefore, is primarily interested in the question of how to build robust systems. This implies that the predictive power of a descriptor is more important than the exact acoustic properties it encodes. In psychology, however, researchers are interested in the insights an audio descriptor can bring to the study of a given perceptual phenomenon. If a descriptor does not add significantly to the overall explanatory power of a model, and if the information it encodes is not transparent, then it should be omitted for the sake of parsimony. These considerations reflect some of the epistemic undercurrents of this topic and explain why studies on timbre psychoacoustics have traditionally used relatively fewer audio descriptors, whereas MIR research on automatic instrument classification used the full gamut of available descriptors.

11.4.2 MPEG-7 Audio Content Description

The multimedia content description interface (Nack and Lindsay 1999; Martínez et al. 2002), also known as MPEG-7, is part of a large effort to standardize multimedia descriptors and descriptor schemes that allow indexing and searching multimedia content, such as pictures, video, audio, and information about how those elements combine in a multimedia context. Unlike the previous MPEG standards that addressed coded representations of audiovisual information, MPEG-7 addresses the representation of information *about* the content, but not the content itself. MPEG-7 began as a scheme for making audiovisual material as searchable as text is today (Nack and Lindsay 1999) and grew to include complex scenarios that employ image processing (such as surveillance) and media conversion, for example, speech to text (Martínez et al. 2002). Within the audio domain, MPEG-7 provides a unified interface for automatic organization of audio from different multimedia sources (i.e., music and film) for applications in sound archiving and classification, and for retrieval, such as music indexing, similarity matching, and MIR (Casey 2001a, b). In addition to traditional timbre description methods that have been applied mainly to isolated musical instrument notes, MPEG-7 also represents noise textures, environmental sounds, music recordings, melodic sequences, vocal utterances (singing and speech), and audio mixtures of the above (Casey 2001b).

MPEG-7 audio comprises text-based description by category labels, also called semantic tags, and quantitative description using audio content descriptors, as explained in Sect. 11.3. Text-based description consists of semantic tags from human annotations (Casey 2001a; Levy and Sandler 2009), whereas audio content descriptors, including descriptors of timbre, are automatically extracted from audio (Lartillot and Toiviainen 2007; Peeters et al. 2011). Audio content descriptors for MPEG-7 include temporal (i.e., the root-mean-squared energy envelope, zero-crossing rate, temporal centroid, and autocorrelation coefficients), spectral (i.e., centroid, flatness, roll-off, and flux), cespstral (i.e., cepstral coefficients and MFCC), perceptual (i.e., sharpness), and specific descriptors (i.e., odd-to-even harmonic energy ratio, harmonic-noise ratio, and attack time).

The semantic tags in text-based descriptions commonly belong to a taxonomy that consists of a number of categories organized into a hierarchical tree used to provide semantic relationships between categories. For example, audio can be categorized into music, speech, or environmental sounds. Each of these categories can be further divided, such as the family of a musical instrument (i.e., brass, woodwinds, strings, and percussion), male or female speech, etc. As the taxonomy gets larger and more fully connected, the utility of the category relationships increases (Casey 2001b). Semantic tags are commonly used in text-based query applications, such as Internet search engines, where text from the query is matched against text from the tags (Casey 2001a). For example, the query "violin" would retrieve sounds tagged with "violin" and possibly also "musical instrument," "strings," etc. Query-by-example applications require audio content descriptors to retrieve sounds in a database that are similar to a target sound provided by the user. In this case, MPEG-7 audio content descriptors are used to compute the similarity with a distance metric such as dynamic time warping for hidden Markov models (Casey 2001b). *Hidden Markov models* are statistical models particularly suited to describe sequences where the probability of the current value depends on the previous value. In fact, Casey points out (also see Sect. 11.3.4) that sound phenomena are dynamic and the descriptors vary in time. In music and speech, this variation carries important information that plays a central role both in perception and in automated tasks. Thus, he proposes to use hidden Markov models in MPEG-7 sound recognition models. Hidden Markov models partition a sound class into a finite number of states, each of which is modeled by a continuous (typically Gaussian) probability distribution. Subsequently, individual sounds are described by their trajectories through this state space, also called *state path*. The state path is an important method of description in the context of MPEG-7 since it describes the evolution of a sound with respect to states that represent events such as onset, sustain, and release (Casey 2001b).

The categorical information in the MPEG-7 tags can be used for automatic classification in which the aim is to automatically assign a class from the taxonomy to audio to which the classifier has not had previous access. Automatic classification involves training statistical models to learn to recognize the class using a descriptor vector as input. Among the most widely used descriptors for automatic audio recognition and classification are representations derived from the power spectrum (Casey 2001a). The raw spectrum is rarely used as input in automatic classification due to the inherent high-dimensionality and redundancy. The typical number of bins of linearly spaced spectra lies between 128 and 1024, whereas probabilistic classifiers, such as hidden Markov models, commonly require low-dimensional representations, preferably fewer than 10 dimensions (Casey 2001b). In MPEG-7, the audio spectrum projection scheme (Casey 2001a; Kim et al. 2004) requires the application of dimensionality reduction techniques, such as principal component analysis or independent component analysis, prior to classification or query-by-example.

Casey concluded that MPEG-7 yielded very good recognizer performance across a broad range of sounds with applications in music genre classification. However, works that compared MFCC with MPEG-7 descriptors in multimedia indexing tasks, such as recognition, retrieval, and classification, found that MFCC outperformed MPEG-7.

Kim et al. (2004) compared the performance of MPEG-7 audio spectrum projection descriptors against MFCC in a video sound track classification task. They used three matrix decomposition algorithms to reduce the dimensionality of the MPEG-7 audio spectrum projection descriptors to 7, 13, and 23 dimensions and compared the resulting approaches with the same number of MFCC. They found that MFCC yielded better performance than MPEG-7 in most cases. They also pointed out that MPEG-7 descriptors are more computationally demanding to extract than MFCC.

Similarly, Deng et al. (2008) compared the performance of traditional descriptors (zero-crossing rate, root-mean-squared energy, spectral centroid, spectral spread, and spectral flux) with MFCC and MPEG-7 on automatic instrument classification tasks. They used several classification algorithms in musical instrument family classification, individual instrument classification, and classification of solo passages. Principal component analysis was used to reduce the dimensionality of MPEG-7 audio spectrum projection descriptors. They concluded that MFCC outperformed MPEG-7 and traditional descriptors when used individually.

Finally, Deng et al. (2008) tested descriptor combinations, such as MFCC with MPEG-7 and MFCC with traditional descriptors, and concluded that the addition of MPEG-7 to MFCC improved classification performance, whereas traditional descriptors plus MFCC yielded the poorest performance. They finally noted that the higher the dimensionality of the descriptors vector, the better the performance; so they tested the classification performance of descriptor combinations followed by dimensionality reduction with principal component analysis and found that the combinations exhibit strong redundancy.

MPEG-7 is a very ambitious international standard that encompasses audio, video, and multimedia description. MPEG-7 audio was developed to have a similar scale of impact on the future of music technology as the MIDI and MPEG-1 Audio Layer III (popularized as the MP3 format) standards have had in the past (Casey 2001b). However, more than 15 years after the introduction of the standard, the world of audio content descriptors still seems anything but standardized—researchers and practitioners continue to develop new approaches that are custom-made for the specific content-description problem at hand.

11.4.3 Computer-Aided Orchestration

Musical orchestration denotes the art of creating instrumental combinations, contrasts, and stratifications (see McAdams, Chap. 8). Initially, orchestration was restricted to the assignment of instruments to the score and, as such, was largely relegated to the background of the compositional process. Progressively, composers started regarding orchestration as an integral part of the compositional process whereby the musical ideas themselves are expressed. Compositional experimentation in orchestration originates from the desire to achieve musically intriguing timbres by means of instrumental combinations. However, orchestration manuals are notoriously empirical because of the difficulty in formalizing knowledge about the timbral result of instrument combinations.

Computer-aided orchestration tools (Carpentier et al. 2010a; Caetano et al. 2019) automate the search for instrument combinations that perceptually approximate a given reference timbre. The aim of computer-aided orchestration is to find a combination of notes from musical instruments that perceptually approximates a given reference sound when played together (Abreu et al. 2016; Caetano et al. 2019). Descriptors of timbre play a key role in the following steps of computer-aided orchestration: (1) timbre description of isolated sounds, (2) timbre description of combinations of musical instrument sounds, and (3) timbre similarity between instrument combinations and the reference sound.

The timbre of both the reference sound and of the isolated musical instrument sounds is represented with a descriptor vector comprising a subset of the traditional descriptors of timbre (Peeters et al. 2011). The extraction of the descriptors is computationally expensive, so the descriptors of the isolated musical instrument sounds are extracted prior to the search for instrument combinations and kept as metadata in a descriptor database. The descriptors for the reference sound are extracted for every new reference used.

Each instrument combination corresponds to a vector of descriptors that captures the timbral result of playing the instruments together. However, the total number of instrument combinations makes it impractical to extract descriptors for each possible combination (Carpentier et al. 2010a). Instead, the descriptor vector of an instrument combination is estimated from the descriptor vectors of the isolated sounds used in the combination (Carpentier et al. 2010b).

The timbral similarity between the reference sound and the instrument combination is estimated as the distance between the corresponding descriptor vectors. Smaller distances indicate a higher degree of timbral similarity (Carpentier et al. 2010a) with the reference, so the instrument combinations with the smallest distances are returned as proposed orchestrations for a given reference sound.

The resulting instrument combinations found to orchestrate a given reference sound will depend on which descriptors are included in the descriptor vector. For example, spectral shape descriptors focus on approximating the distribution of spectral energy of the reference sound. Carpentier et al. (2010a) proposed using the normalized harmonic energy, global noisiness, attack time, spectral flatness, roughness, frequency and amplitude of the energy modulation, and frequency and amplitude of the modulation of the fundamental frequency. Additionally, they added the following descriptors not related to timbre: fundamental frequency and total energy. Caetano et al. (2019) used the frequency and amplitude of the spectral peaks, spectral centroid, spectral spread, and also fundamental frequency, loudness, and root-mean-squared energy.

The type of reference sound to be orchestrated also plays a fundamental role in the instrument combinations found by computer-aided orchestration algorithms. For example, if the composer uses a clarinet sound as reference (and the musical instrument sound database contains clarinet sounds), the composer should naturally expect an isolated clarinet note to be the closest instrument combination found (unless the composer imposes constraints to the search such as returning instrument combinations without clarinet sounds or with at least three different instruments). Aesthetically interesting results can be achieved by choosing a reference sound that belongs to a different abstract category than musical instruments, such as environmental sounds or vocal utterances, because these references usually result in complex instrument combinations. To hear examples of orchestrations using different types of reference sounds, go to the sound files "car_horn.mp3", "carnatic.mp3", "choir.mp3", and wind_harp.mp3". Each sound file consists of the reference sound followed by four proposed orchestrations from Caetano et al. (2019).

Perceptually, two important phenomena contribute to attaining aesthetically interesting orchestrations: *timbre blends* and *sensory proximity*. Timbre blends occur when the timbre of the different instruments used in the combination fuse into a single percept (see McAdams, Chap. 8). The categorical distinction between the musical instruments must disappear so the sensory attributes of the combination emerge as a new timbre. Computer-aided orchestration algorithms find instrument combinations whose sensory attributes approximate those of the reference sound to evoke abstract auditory experiences. Audio descriptors of timbre play a key role in determining the timbre similarity between the instrument combinations and the reference sound (Siedenburg et al. 2016a). Traditionally, timbre similarity approaches compare time-averaged descriptor vectors from different musical instrument sounds, neglecting temporal variations (Esling and Agon 2013). While this is consistent with static timbre spaces, dynamic representations, such as the one shown in Fig. 11.7, require the use of time series of descriptors.

Computer-aided orchestration exemplifies the benefit of incorporating temporal information into timbre similarity. The static timbre similarity measure is appropriate when orchestrating reference sounds that are relatively stable (Carpentier et al. 2010a; Abreu et al. 2016). However, matching targets with dynamic variations, such as an elephant trumpeting, requires a time-series method that takes temporal variations of descriptors into consideration. Esling and Agon (2013) proposed a multi-objective time series-matching algorithm capable of coping with the temporal and multidimensional nature of timbre. The multi-objective time series-matching algorithm adopts a multi-dimensional measure of similarity that simultaneously optimizes the temporal evolution of multiple spectral properties and returns a set of efficient solutions rather than a single best solution.

11.4.4 Musical Instrument Sound Morphing

The aim of sound morphing is to synthesize sounds that gradually blur the categorical distinction between the sounds being morphed by blending their sensory attributes (Caetano and Rodet 2013). Therefore, sound morphing techniques allow synthesizing sounds with intermediate timbral qualities by interpolating the sounds being morphed. Fig. 11.8 shows a striking example of image morphing to illustrate sound morphing with a visual analogy. The morph is determined by a single parameter α that varies between 0 and 1. Only the source sound S is heard when $\alpha = 0$, whereas only the target sound T is heard when $\alpha = 1$. Intermediate values of α should correspond to perceptually intermediate sounds. However, simple morphing techniques seldom satisfy this perceptual requirement (Caetano and Rodet 2013). Sound morphing typically comprises the following steps: (1) modeling, (2) interpolation, and (3) resynthesis.

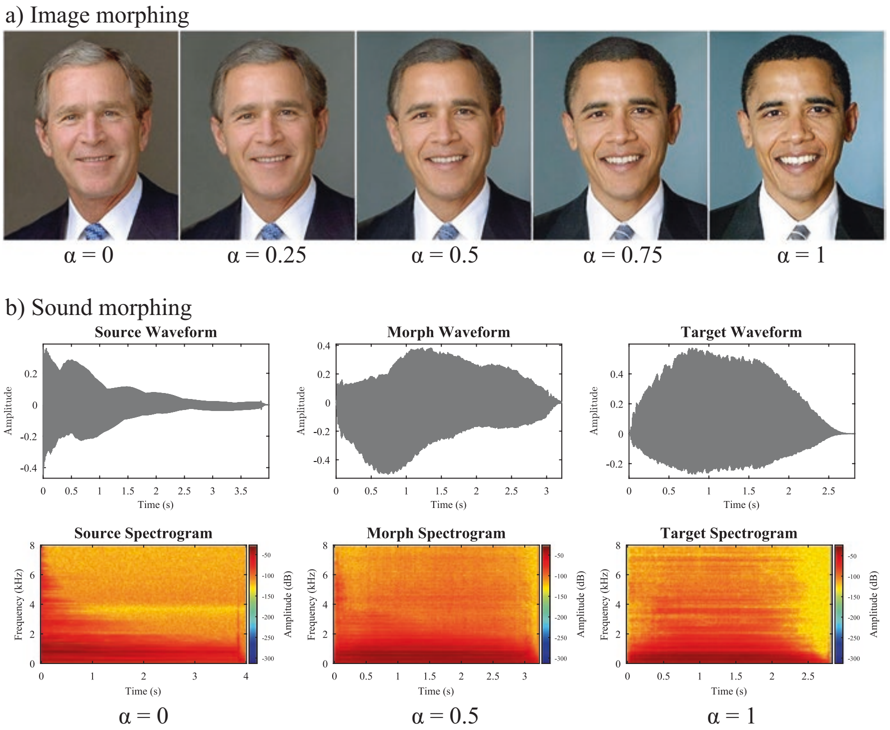

Fig. 11.8 Illustration of morphing for images and sounds: (a) face morphing; (b) musical instrument sound morphing. The source sound is the C#3 note played *forte* on a harpsichord and the target sound is the same note played *forte* on a tuba. The figure shows the *morphing factor* **α** below each corresponding panel. To hear the sounds, go to the sound file "harpsichord_tuba_morph.mp3". The image in (a) is currently publicly available at https://paulbakaus.com/wp-content/uploads/2009/10/bush-obama-morphing.jpg

The sounds being morphed (*S* and *T*) are modeled (e.g., with the sinusoidal model or the excitation-filter model) to obtain a parametric representation of *S* and *T*. For example, the parameters of the sinusoidal model are the frequencies and the amplitudes of the time-varying sinusoids that represent the partials of *S* and *T*. The parameters of the spectral envelope represent the filter component of the excitation-filter model. Cepstral coefficients and LPC are common representations of the filter in the excitation-filter model (Caetano and Rodet 2013).

The parameters of the morphed sound are obtained via linear interpolation between the parameters of *S* and *T*, for example, interpolation of the amplitudes and frequencies of the sinusoidal model or interpolation of the cepstral coefficients representing the spectral envelope of the excitation-filter model.

Finally, the morphed sound is resynthesized from the interpolated parameters. Perceptually, the model parameters play a crucial role in the final result, depending on the information captured by the model. For example, morphing with the sinusoidal model will result in intermediate amplitudes and frequencies (the model parameters), whereas morphing with the excitation-filter model will result in intermediate spectral envelopes.

The importance of the parametric representation is twofold: resynthesis and transformation. A parametric model should allow resynthesizing a sound that is perceptually very similar to the original sound from the model parameters alone. Sound transformations are achieved via manipulation of the model parameters followed by resynthesis, resulting in a sound that is perceptually different from the original sound. Striking transformations can be achieved by careful manipulation of model parameters depending on what information they represent. For example, the frequencies of the sinusoidal model can be manipulated to obtain a pitch transposition. Sound morphing is the result of parameter interpolation. However, most morphing techniques in the literature interpolate the parameters of the model used to represent the sounds regardless of the perceptual impact of doing so. Consequently, the morph is intermediate in the space of parameters rather than perceptually intermediate.

Caetano and Rodet (2013) used descriptors of timbre to guide musical instrument morphing toward more gradual transformations. They developed a sophisticated morphing technique based on a hybrid excitation-filter model where the filter is represented with spectral envelopes and the excitation has a sinusoidal component accounting for the partials and a residual component accounting for transients and noise missed by the sinusoids. Caetano and Rodet (2013) investigated the result of interpolating several representations of the spectral envelope: the spectral envelope curve, cepstral coefficients, LPC, reflection coefficients, and line spectral frequencies. Both reflection coefficients and line spectral frequencies arise from an interconnected tube model of the human vocal tract. Reflection coefficients represent the fraction of energy reflected at each section of the model, whereas line spectral frequencies represent the resonance conditions that describe the vocal tract being fully open or fully closed at the glottis (McLoughlin 2008).

Caetano and Rodet (2013) were interested in measuring the linearity of the morphing transformation with the different spectral envelope representations. They varied α linearly between 0 and 1 for each spectral envelope representation and recorded the corresponding variation of spectral shape descriptors (spectral centroid, spectral spread, spectral skewness, and spectral kurtosis). They found that linear interpolation of line spectral frequencies led to the most linear variation of spectral shape descriptors. Next, they performed a listening test to evaluate the perceptual linearity of the morphs with their hybrid excitation-filter model and the sinusoidal model. The listening test confirmed that the hybrid excitation-filter model resulted in morphs that were perceived as more perceptually linear than the sinusoidal model. Fig. 11.8 shows an example of musical instrument sound morphing from Caetano and Rodet (2013). To hear the sounds used in Fig. 11.8, go to the sound file "harpsichord_tuba_morph.mp3".

Perceptually, sound morphing can be viewed as an auditory illusion that is inherently intertwined with timbre because morphing manipulates both the sensory and the categorical perceptions of the sounds being morphed. For the sake of simplicity, the following examples will consider musical instruments and timbre spaces. In theory, sound morphing can break the categorical perception of musical instrument timbre. For example, when S and T are from different musical instruments, setting $\alpha = 0.5$ would produce a morph that theoretically resembles the sound of a hybrid musical instrument. Additionally, sound morphing can be used to create a *sonic continuum*. Timbre spaces are inherently sparse, with musical instrument sounds occupying specific points in an otherwise void space. Morphing musical instrument sounds can theoretically fill the gaps and create continuous timbre spaces by connecting musical instruments with intermediate sounds that no acoustical instrument can produce.

11.5 Summary

This chapter introduced the acoustic modeling of timbre via audio content descriptors. Sections were organized around the descriptor extraction process, covering important topics from general audio representations used to extract timbre descriptors to applications of these descriptors in psychology, sound synthesis, and music information retrieval. Audio content descriptors have played an important role in understanding the psychoacoustics of timbre, have become part of the industry standard MPEG-7 for audio content description, and play crucial roles for current developments of techniques such as computer-aided orchestration and musical instrument sound morphing. In these applications, audio descriptors help extract properties from the audio signal that are often of perceptual relevance and much more specific when compared to the general audio representations from which they are computed. At the same time, the audio descriptors described in this chapter are versatile enough to be valuable across a variety of different timbre-related audio processing tasks.

Audio descriptors could play a pivotal role in future research into timbre perception and sound processing in myriad ways. Section 11.4.1 outlined the ways in which the perception of timbral brightness has been modeled on acoustic grounds using audio descriptors. However, a model of timbre brightness perception that clearly delineates the acoustic ingredients of this important aspect of timbre perception has yet to be constructed and evaluated. Such a model would need to account for a variety of experimental phenomena (see McAdams, Chap. 2) across a large set of sounds. Section 11.4.3 and 11.4.4 summarized the role of audio descriptors in computer-aided orchestration and sound morphing. Here audio descriptors act as a perceptual proxy to allow synthesizing sounds with predefined perceptual characteristics. Adaptive processing (Verfaille et al. 2006) and content-based transformations (Amatriain et al. 2003) use audio descriptors to address the highly nonlinear connection between the audio and sound perception. However, the fundamental problem of synthesizing a waveform that matches a desired perceptual result remains a challenge.

Currently, the status of various approaches to audio content description is at a crossroads. The rise of machine learning architectures, such as deep neural networks, renders traditional audio descriptors obsolete in tasks such as musical instrument identification, environmental scene classification, or speaker recognition. Traditional audio descriptor-based classification architectures require two steps prior to learning per se: descriptor extraction followed by either descriptor selection or dimensionality reduction (see Sect. 11.3.4). One problem of these architectures is that they often fail to capture the highly nonlinear relationships commonly found in complex classification tasks. Deep neural networks are feed-forward artificial neural networks with several layers of hidden units between inputs and outputs (Hinton et al. 2012). The depth of the network provides sufficient flexibility to represent the nonlinearities critical to a given task such that deep neural networks jointly learn the descriptors and the classifier (Takahashi et al. 2018).

However, the main challenge of deep learning architectures lies in their application in timbre acoustics, perception, and cognition. Kell et al. (2018) made a significant contribution when they presented a deep neural network optimized for both speech and music recognition tasks. The deep neural network performed as well as humans, exhibited error patterns that resembled those of humans, and outperformed a linear spectrotemporal filter model of auditory cortex in the prediction of fMRI voxel responses. Moreover, the trained network replicated aspects of human cortical organization and provided evidence of hierarchical organization within the auditory cortex, with intermediate and deep layers best predicting primary and nonprimary auditory cortical responses, respectively. Nonetheless, prediction is not identical to understanding. Even though a good model should predict future data, a model needs to be transparent in order to allow for proper theory building. Future work in this direction will be able to draw insightful connections between the pattern of oscillations carried by sound waves and the timbre that listeners extract from these waves.

Compliance with Ethics Requirements Marcelo Caetano declares that he has no conflict of interest.

Charalampos Saitis declares that he has no conflict of interest.

Kai Siedenburg declares that he has no conflict of interest.

References

Abreu J, Caetano M, Penha R (2016) Computer-aided musical orchestration using an artificial immune system. In: Johnson C, Ciesielski V, Correia J, Machado P (eds) Evolutionary and biologically inspired music, sound, art and design, lecture notes in computer science, vol 9596. Springer, Heidelberg, pp 1–16

Almeida A, Schubert E, Smith J, Wolfe J (2017) Brightness scaling of periodic tones. Atten Percept Psychophys 79(7):1892–1896

Amatriain X, Bonada J, Loscos À et al (2003) Content-based transformations. J New Music Res 32(1):95–114

Aucouturier J-J, Defreville B, Pachet F (2007) The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. J Acoust Soc Am. https://doi.org/10.1121/1.2750160

Barthet M, Depalle P, Kronland-Martinet R, Ystad S (2010) Acoustical correlates of timbre and expressiveness in clarinet performance. Music Percept 28(2):135–153

Bogert BP, Healy MJR, Tukey JW (1963) The quefrency analysis of time series for echoes: cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking. In: Rosenblatt M (ed) Time series analysis. Wiley, New York, pp 209–243

Brown JC (1991) Calculation of a constant Q spectral transform. J Acoust Soc Am 89(1):425–434

Brown JC (1999) Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. J Acoust Soc Am 105(3). https://doi.org/10.1121/1.426728

Brown JC, Houix O, McAdams S (2001) Feature dependence in the automatic identification of musical woodwind instruments. J Acoust Soc Am 109(3):1064–1072. https://doi.org/10.1121/1.1342075

Brown JC, Puckette MS (1992) An efficient algorithm for the calculation of a constant q transform. J Acoust Soc Am 92(5):2698–2701

Burred JJ, Röbel A (2010) A segmental spectro-temporal model of musical timbre. In: Zotter F (ed) Proceedings of the 13th international conference on digital audio effects (DAFx-10). IEM, Graz

Burred JJ, Röbel A, Sikora T (2010) Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds. IEEE Trans Audio Speech Lang Proc 18(3):663–674

Caclin A, McAdams S, Smith BK, Winsberg S (2005) Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones. J Acoust Soc Am 118:471–482

Caetano MF, Burred JJ, Rodet X (2010) Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues. In: Zoter F (ed) Proceedings of the 13th international conference on digital audio effects (DAFx-10). IEM, Graz

Caetano M, Rodet X (2013) Musical instrument sound morphing guided by perceptually motivated features. IEEE Trans Audio Speech Lang Proc 21(8):1666–1675

Caetano M, Zacharakis A, Barbancho I, Tardón LJ (2019) Leveraging diversity in computer-aided musical orchestration with an artificial immune system for multi-modal optimization. Swarm Evol Comput. https://doi.org/10.1016/j.swevo.2018.12.010

Carpentier G, Assayag G, Saint-James E (2010a) Solving the musical orchestration problem using multiobjective constrained optimization with a genetic local search approach. J Heuristics 16(5):681–714. https://doi.org/10.1007/s10732-009-9113-7

Carpentier G, Tardieu D, Harvey J et al (2010b) Predicting timbre features of instrument sound combinations: application to automatic orchestration. J New Mus Res 39(1):47–61

Casey M (2001a) MPEG-7 sound-recognition tools. IEEE Trans Circ Sys Video Tech 11(6):737–747

Casey M (2001b) General sound classification and similarity in MPEG-7. Organized Sound 6(2):153–164

Casey MA, Veltkamp R, Goto M et al (2008) Content-based music information retrieval: current directions and future challenges. Proc IEEE 96(4):668–696

Childers DG, Skinner DP, Kemerait RC (1977) The cepstrum: a guide to processing. Proc IEEE 65(10):1428–1443

Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28(4):357–366

Deng JD, Simmermacher C, Cranefield S (2008) A study on feature analysis for musical instrument classification. IEEE Trans Syst Man Cybern B Cybern 38(2):429–438

De Poli G, Prandoni P (1997) Sonological models for timbre characterization. J New Mus Res 26(2):170–197

Dolson M (1986) The phase vocoder: a tutorial. Comp Music J 10(4):14–27. https://doi.org/10.2307/3680093

Esling P, Agon C (2013) Multiobjective time series matching for audio classification and retrieval. IEEE Trans Audio Speech Lang Proc 21(10):2057–2072

Fletcher NH (1999) The nonlinear physics of musical instruments. Rep Prog Phys 62(5):723–764

Giordano BL, McAdams S, Zatorre RJ et al (2012) Abstract encoding of auditory objects in cortical activity patterns. Cereb Cortex 23(9):2025–2037

Glasberg BR, Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. Hear Res 47:103–138

Grey JM (1977) Multidimensional perceptual scaling of musical timbres. J Acoust Soc Am 61(5). https://doi.org/10.1121/1.381428

Grey JM, Gordon JW (1978) Perceptual effects of spectral modifications on musical timbres. J Acoust Soc Am 63(5):1493–1500

Hajda J (2007) The effect of dynamic acoustical features on musical timbre. In: Beauchamp JW (ed) Analysis, synthesis, and perception of musical sounds. Springer, New York, pp 250–271

Handel S (1995) Timbre perception and auditory object identification. In: Moore BCJ (ed) Hearing, Handbook of perception and cognition, 2nd edn. Academic Press, San Diego, pp 425–461

Harris FJ (1978) On the use of windows for harmonic analysis with the discrete Fourier transform. Proc IEEE 66(1):51–83

Hartmann WM (1996) Pitch, periodicity, and auditory organization. J Acoust Soc Am 100(6):3491–3502

Herrera-Boyer P, Peeters G, Dubnov S (2003) Automatic classification of musical instrument sounds. J New Music Res 32(1):3–21

Hinton G, Deng L, Yu D et al (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig Proc Mag 29(6):82–97

Holighaus N, Dörfler M, Velasco GA, Grill T (2013) A framework for invertible, real-time constant-Q transforms. IEEE Trans Audio Speech Lang Proc 21(4):775–785

Horner AB, Beauchamp JW, So RH (2011) Evaluation of Mel-band and MFCC-based error metrics for correspondence to discrimination of spectrally altered musical instrument sounds. J Audio Eng Soc 59(5):290–303

Huq A, Bello JP, Rowe R (2010) Automated music emotion recognition: a systematic evaluation. J New Mus Res 39(3):227–244

Irino T, Patterson RD (1997) A time-domain, level-dependent auditory filter: the gammachirp. J Acoust Soc Am 101:412–419

Jaffe DA (1987a) Spectrum analysis tutorial, part 1: the discrete Fourier transform. Comp Music J 11(2):9–24

Jaffe DA (1987b) Spectrum analysis tutorial, part 2: properties and applications of the discrete Fourier transform. Comp Music J 11(3):17–35

Kell AJE, Yamins DLK, Shook EN et al (2018) A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98(3):630–644. https://doi.org/10.1016/j.neuron.2018.03.044

Kim HG, Burred JJ, Sikora T (2004) How efficient is MPEG-7 for general sound recognition? Paper presented at the 25th international Audio Engineering Society conference: metadata for audio. London, 17–19 June 2004

Krimphoff J, Mcadams S, Winsberg S (1994) Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique (Characterization of the timbre of complex sounds II Acoustic analysis and psychophysical quantification) J de Physique (J Phys) IV(C5):625–628

Lartillot O, Toiviainen P (2007) A Matlab toolbox for musical feature extraction from audio. In: Marchand S (ed) Proceedings of the 10th international conference on digital audio effects (DAFx-07). Université de Bordeaux, Bordeaux, p 237–244

Levy M, Sandler M (2009) Music information retrieval using social tags and audio. IEEE Trans Multimedia 11(3):383–395

Lyon FL (2017) Human and machine hearing: extracting meaning from sound. Cambridge University Press, Cambridge

McLoughlin IV (2008) Review: line spectral pairs. Sig Proc 88(3):448–467

Makhoul J (1975) Linear prediction: a tutorial review. Proc IEEE 63(4):561–580

Martínez JM, Koenen R, Pereira F (2002) MPEG-7: the generic multimedia content description standard, part 1. IEEE MultiMedia 9(2):78–87

Marozeau J, de Cheveigné A (2007) The effect of fundamental frequency on the brightness dimension of timbre. J Acoust Soc Am 121(1):383–387

Martínez J, Perez H, Escamilla E, Suzuki MM (2012). Speaker recognition using mel frequency cepstral coefficients (MFCC) and vector quantization (VQ) techniques. In: Sánchez PB (ed) Proceedings of the 22nd international conference on electrical communications and computers. IEEE, Piscataway, p 248–251

McAdams S, Douglas C, Vempala NN (2017) Perception and modeling of affective qualities of musical instrument sounds across pitch registers. Front Psychol. https://doi.org/10.3389/fpsyg.2017.00153

McAdams S, Winsberg S, Donnadieu S et al (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res 58(3):177–192

McAulay R, Quatieri T (1986) Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans Acoust Speech Sig Proc 34(4):744–754

McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. Nat Neurosci 16:493–498

Nack F, Lindsay AT (1999) Everything you wanted to know about MPEG-7: part 2. IEEE MultiMedia 6(4):64–73

Ogg M, Slevc LR, Idsardi WJ (2017) The time course of sound category identification: insights from acoustic features. J Acoust Soc Am 142(6):3459–3473

On CK, Pandiyan PM, Yaacob S, Saudi A (2006). Mel-frequency cepstral coefficient analysis in speech recognition. Paper presented at the 2006 international conference on computing & informatics (ICOCI 2006). Kuala Lumpur, 6–8 June 2006

Patterson RD, Robinson K et al (1992) Complex sounds and auditory images. In: Cazals Y, Demany L, Horner K (eds) Auditory physiology and perception. Pergamon Press, Oxford, pp 429–446

Peeters G, Giordano BL, Susini P et al (2011) The timbre toolbox: audio descriptors of musical signals. J Acoust Soc Am 130:2902–2916. https://doi.org/10.1121/1.3642604

Pollard HF, Jansson EV (1982) A tristimulus method for the specification of musical timbre. Acta Acust united Ac 51(3):162–171

Portnoff M (1980) Time-frequency representation of digital signals and systems based on short-time Fourier analysis. IEEE Trans Acoust Speech Sig Proc 28(1):55–69

Regnier L, Peeters G (2009) Singing voice detection in music tracks using direct voice vibrato detection. In: Chen LG, Glass JR (eds) Proceedings of the 2009 IEEE international conference on acoustics, speech and signal processing, Taipei, April 2009. IEEE, Piscataway, p 1685–1688

Rigaud F, David B (2013) A parametric model and estimation techniques for the inharmonicity and tuning of the piano. J Acoust Soc Am 133(5):3107–3118. https://doi.org/10.1121/1.4799806

Saitis C, Giordano BL, Fritz C, Scavone GP (2012) Perceptual evaluation of violins: a quantitative analysis of preference judgements by experienced players. J Acoust Soc Am 132:4002–4012

Schubert E, Wolfe J (2006) Does timbral brightness scale with frequency and spectral centroid? Acta Acust united Ac 92(5):820–825

Siedenburg K, Fujinaga I, McAdams S (2016a) A comparison of approaches to timbre descriptors in music information retrieval and music psychology. J New Music Res 45(1):27–41

Siedenburg K, Jones-Mollerup K, McAdams S (2016b) Acoustic and categorical dissimilarity of musical timbre: evidence from asymmetries between acoustic and chimeric sounds. Front Psychol 6(1977)

Siedenburg K, McAdams S (2017) Four distinctions for the auditory "wastebasket" of timbre. Front Psychol 8(1747)

Slawson W (1985) Sound color. University of California Press, Berkeley

Stevens SS, Volkman J, Newman E (1937) A scale for the measurement of the psychological magnitude of pitch. J Acoust Soc Am 8(3):185–190

Takahashi N, Gygli M, Van Gool L (2018) AENet: learning deep audio features for video analysis. IEEE Trans Multimedia 20(3):513–524

Terasawa H, Slaney M, Berger J (2005) The thirteen colors of timbre. In: proceedings of the 2005 IEEE workshop on applications of signal processing to audio and acoustics, new Paltz, October 2005. IEEE, Piscataway, p 323–326

Verfaille V, Zolzer U, Arfib D (2006) Adaptive digital audio effects (a-DAFx): a new class of sound transformations. IEEE Trans Audio Speech Lang Proc 14(5):1817–1831

Zwicker E (1961) Subdivision of the audible frequency range into critical bands (Frequenzgruppen). J Acoust Soc Am 33:248–248

Zwicker E, Fastl H (1990) Psychoacoustics: facts and models. Springer, Berlin

Chapter 12
Modulation Representations for Speech and Music

Mounya Elhilali

Abstract The concept of modulation has been ubiquitously linked to the notion of timbre. Modulation describes the variations of an acoustic signal (both spectrally and temporally) that shape how the acoustic energy fluctuates as the signal evolves over time. These fluctuations are largely shaped by the physics of a sound source or acoustic event and, as such, are inextricably reflective of the sound identity or its timbre. How one extracts these variations or modulations remains an open research question. The manifestation of signal variations not only spans the time and frequency axes but also bridges various resolutions in the joint spectrotemporal space. The additional variations driven by linguistic and musical constructs (e.g., semantics, harmony) further compound the complexity of the spectrotemporal space. This chapter examines common techniques that are used to explore the modulation space in such signals, which include signal processing, psychophysics, and neurophysiology. The perceptual and neural interpretations of modulation representations are discussed in the context of biological encoding of sounds in the central auditory system and the psychophysical manifestations of these cues. This chapter enumerates various representations of modulations, including the signal envelope, the modulation spectrum, and spectrotemporal receptive fields. The review also examines the effectiveness of these representations for understanding how sound modulations convey information to the listener about the timbre of a sound and, ultimately, how sound modulations shape the complex perceptual experience evoked by everyday sounds such as speech and music.

Keywords Auditory cortex · Modulation spectrum · Musical signal
Spectrotemporal modulations · Spectrotemporal receptive fields · Speech

M. Elhilali (✉)
Laboratory for Computational Audio Perception, Center for Speech and Language Processing, Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA
e-mail: mounya@jhu.edu

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_12

12.1 Introduction

If one asks a telecommunication engineer what is “modulation”, the answer is rather simple: It is the process of multiplexing two signals: a signal that can carry information and can be physically transmitted over a communication channel (the *carrier signal*, typically a quickly varying wave) with a signal that contains the information or the message to be transmitted or broadcasted (the *modulation* or *data signal*, typically a slowly varying envelope) (Freeman 2004). This characterization provides a formal account of modulation but fails to capture the nuances of multiplexing two signals that get rather complicated depending on the domain under study. This definition presumes a priori knowledge of the identity, attributes, and behavior of such signals, which is only possible in specific applications (e.g., on/off keying—OOF—used to transmit binary 0/1 codes over a sinusoidal carrier that can be decoded directly from the signal amplitude).

On the flip side, defining modulation as a multiplexing operation is rather ineffective when it comes to the inverse problem: demodulating a signal in order to identify its modulator and carrier components. If one does not have specific constraints on these signal components, it is not trivial to untangle them because many (possibly infinite) solutions are conceivable. How one judges which solution is a reasonable one is again domain and signal specific. As such, the modulation/demodulation problem is ill-posed (Turner and Sahani 2011) but is still fundamental to understanding the information-bearing components of signals.

In the case of complex audio signals (speech, music, natural, or communication sounds), getting a clear idea of the identity of the message and carrier components remains one of the holy grails of research on the physical and perceptual underpinnings of sound. Interest in modulations of an audio signal aims to pinpoint the information-bearing components of these signals, especially given the redundant nature of the waveforms that can emanate from both mechanical (e.g., instrument, vocal tract) or electrical (e.g., computer generated) sound sources.

The problem is particularly compounded because complex audio signals, such as speech and music, contain information and modulations at multiple time scales and across various spectral constructs. In the case of speech, there is an extensive body of work dating back to the early twentieth century that explored the span and dynamics of the speech envelope. The argument that the slow envelope is the chief carrier of phonetic information in speech is quite old. In the 1930’s, Dudley advocated that the dynamics of signal envelopes are important for describing linguistic information in speech (Dudley 1939, 1940). In his view, the vocal tract is a sluggish system that slowly changes shape, with low syllabic frequencies up to 10 Hz, giving rise to varying modulating envelopes that contribute most to the intelligibility of speech.

Building on this work, numerous studies have shown that speech intelligibility is well maintained after temporal envelopes are lowpass filtered or degraded, with a critical range between 5–15 Hz that spans the range of phonemic and syllabic rates in natural speech (Greenberg 2004). Still, the modulation spectrum profile of speech is a complex one and reveals that the speech envelope contains energy of the order of a few to tens or hundreds of Hertz. This profile highlights key energy fluctuations in speech signals, ranging from hundreds of milliseconds (of the order of multiple syllables or words) to tens of milliseconds (typically spanning subsyllabic and phonemic segments) (Rosen 1992; Divenyi et al. 2006). The complexity of speech signals includes the multiplexed information across various time scales but also variations across frequency bands and in the phase relationships across bands (Pickett 1999).

In the case of music signals, a similar picture emerges spanning multiple time scales, frequency bands, and spectral profiles. The information-bearing components of a musical signal, be it an isolated note or a full orchestral piece, appear to multiplex across a complex construct of spectrotemporal dimensions. Much like speech, music signals have melodic, harmonic, and rhythmic structures that intertwine into intricate patterns (both in time and frequency) to convey the complex acoustic experience of music perception. Recent advances in computing power, signal processing techniques, and increased availability of digitized audio material have led to fairly sophisticated analysis tools to study various aspects of regularity in music, such as rhythm, melody, harmony, or timbre (Müller 2015; Meredith 2016).

Despite the intricate nature of spectrotemporal regularities in both speech and music, they share fundamental attributes reflected in their decomposition into alphabetic tokens (phonemes, syllables, word, notes, chords), assembly of sequences of events (accents, grouping, words, phrases), and rhythmic structure (time, stress), all interleaved with specific spectral patterns that reflect the sound sources (instrument, oral cavity), production style, and contextual attributes. The correlates of these regularities can be gleaned from examining the modulation patterns in the signal at multiple time scales and granularities. This chapter reviews common techniques used to represent modulations in speech and music signals and their implications for understanding the information-bearing components in these signals. Section 12.2 reviews signal processing tools commonly used to represent modulations: fundamental time-frequency representations (Sect. 12.2.1), spectrotemporal modulation profiles (Sect 12.2.2), and temporal modulation spectra (Sect 12.2.3). Sect. 12.2.4 delves into representations that are unique to speech and music signals and considers constraints imposed by the physics of the vocal tract and controlled sound production through most musical instruments. Section 12.3 offers insights into the neurophysiological interpretation of modulations, particularly encoding of the spectrotemporal signal envelope along the auditory pathway. Section 12.4 reviews key findings in psychophysical and physiological research into the role of modulation in speech and music perception. Section 12.5 provides a summary of the main ideas in the text along with perspectives on future research directions.

12.2 Representation of Modulations

12.2.1 The Time-Frequency Representation

A complete description of the information content of speech and music signals is not possible. However, one can derive a number of low-level empirical descriptors that reveal a lot about the structure of these signals. Common ways to explore the nature of these signals involve analysis of the acoustic waveform as well as its frequency content. A time-frequency profile, typically obtained via short-time Fourier transform, wavelet, or filterbank analysis (Caetano, Saitis, and Siedenburg, Chap. 11), best displays the variations of energy as the signal evolves over time. Fig. 12.1A depicts the time-frequency representation of a speech utterance produced by a male speaker saying /we think differently/. Immediately emerging from this spectrographic view of speech is the fact that the temporal envelope varies slowly over the course of tens to hundreds of milliseconds. In fact, one can easily discern the volleys of activity across frequency channels, occurring at a rate of 5–7 peaks per second, commensurate with phonemic and syllabic contours of the speech utterance. The right subpanel in Fig. 12.1A highlights a cross-section of this spectrogram around 450 Hz, which represents the half-wave rectified output of the auditory filter centered about that spectral region. The time waveform clearly shows an overall fluctuating pattern around 6 Hz, which closely follows segmental and syllabic landmarks of the speech signal (Poeppel et al. 2008). A similar structure emerges spectrally with frequency profiles that are largely coarse. The energy distribution across frequency channels appears to mostly delineate harmonic and formant peaks (bottom-left subpanel in Fig. 12.1A).

In parallel, Fig. 12.1B illustrates an example of the time-frequency spectrogram of the finale of Tchaikovsky's violin concerto in D major, Op. 35. The time-frequency spectrogram highlights the exuberant energy in this piece with a very dynamic temporal profile reflective of the vigorous nature of this finale. The clear steady tones typical of bowed string instruments are also clearly visible throughout the passage, with the spectral profile showing the clear harmonic nuances of the solo violin performance. Still, the rich energy of this final movement of the concert is not readily discernable from the spectrogram view only. The cross-section of this spectrogram along time emphasizes the nested dynamics over the course of a 6 s period. The soft onset signature of the violin is not very evident due to the multiscale rhythmic modulations in this extravagantly energetic piece with discernable Russian rhythmic undertones (Sadie 2001). The temporal envelope clearly shows a fast-paced profile modulated by a much slower rhythmic profile varying at rate of 1–3 peaks/s. The spectral cross-section shown in the bottom-left panel in Fig. 12.1B takes a closer look at the frequency profile of the signal around 2.3 s. The characteristic profile of a violin note clearly emerges with the overall envelope highlighting the resonance of the violin body with three obvious peaks (Katz 2006). Within the broad peaks, one can glimpse details of the spectral structure imposed by the mechanical constraints of the violin along with the unambiguous harmonic structure of the note.

Fig. 12.1 (Continued) as a function of time; a *frequency cross-section* of the spectrogram around 250 ms as a function of log-frequency; and the *two-dimensional Fourier transform* (*2D FFT*) of the time-frequency spectrograms that yields the modulation power spectrum of the signal. The figure was interpolated using linear interpolation and compressed to a power of 2.5 to obtain better color contrast (for display purposes only). (**B**) The spectrotemporal details of the finale of Tchaikovsky's violin concerto in D major, Opus 35, using similar processing steps as in panel **A**

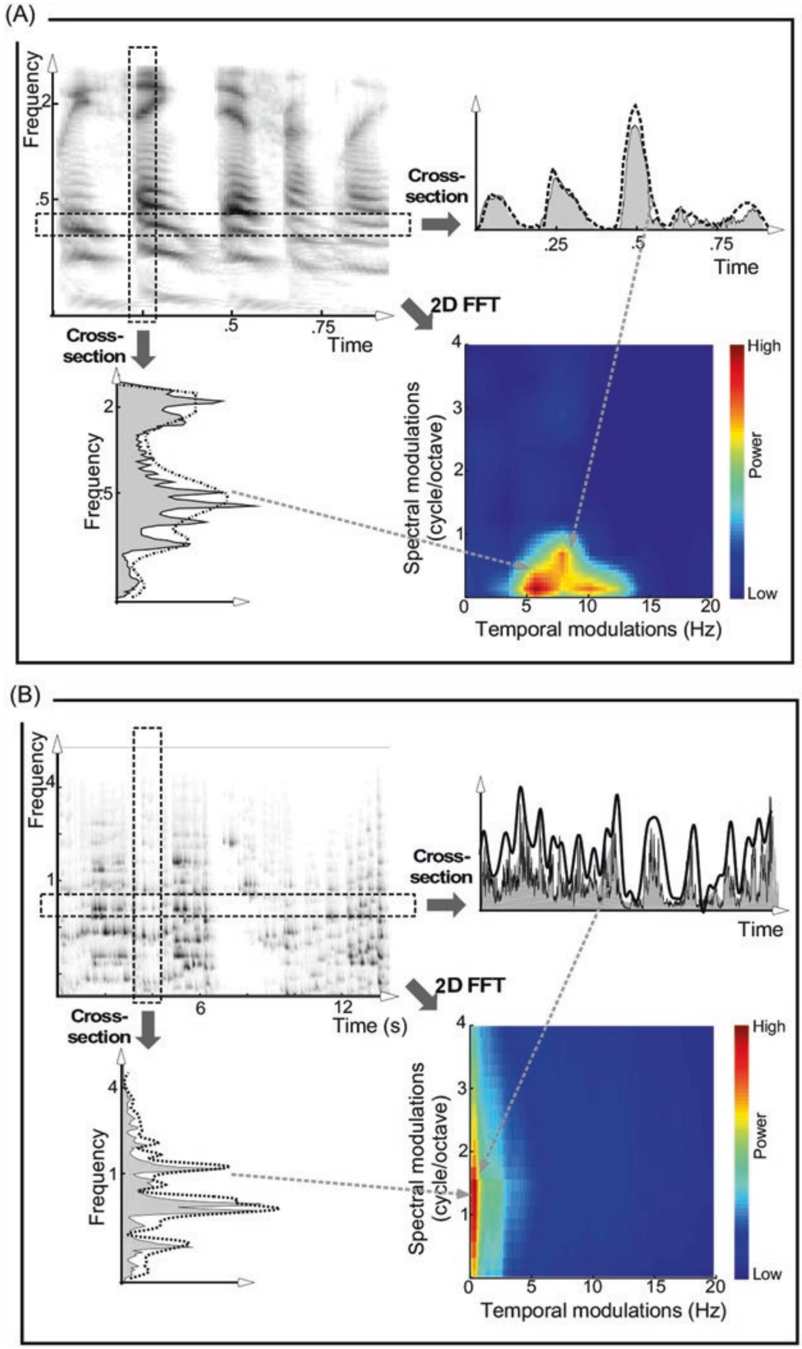

Fig. 12.1 Spectrotemporal details of speech and music. (**A**) The *time-frequency spectrogram* of a male utterance saying /we think differently/ over a time span of 1 s and frequency range of 5 octaves (note the log frequency axis); a *temporal cross-section* of the spectrogram around 450 Hz

12.2.2 The Spectrotemporal Modulation Profile

A better illustration of these spectrotemporal modulation details can be achieved in the spectral/Fourier domain, obtained by performing a two-dimensional Fourier transform of the time-frequency spectrogram (Fig. 12.1A, B, lower right panels). This operation estimates the power distribution of both spectral and temporal components over the chosen time and frequency spans and yields the *modulation spectrum* of the signal (Singh and Theunissen 2003). The modulation spectrum is an account of the distribution of time-frequency correlations of adjacent and far-away elements in the signal and, hence, is an estimate of the degree and dynamics of signal fluctuations along time and frequency axes. Immediately worth noting in this modulation spectrum is that the energy in the Fourier domain is mostly concentrated in a highly localized region of the modulation space.

For a speech signal (Fig. 12.1A), the modulation spectrum highlights what was already seen in the unidimensional profiles. For instance, the temporal envelope induces a strong activation peak between 5 and 7 Hz, while the spectral modulations reveal discernable energy at a harmonic rate (i.e., distance between harmonic peaks) or coarser (i.e., distance between formant peaks), which appear as strong activity around 1 cycle/octave and below. The modulation spectrum energy for the music signal (Fig. 12.1B) also accentuates the modulation patterns observed in the cross-sections of the spectrogram. A strong activation pattern around 1 cycle/octave clearly highlights the crisp harmonic peaks of a violin sound, while the temporal modulations show a distributed energy that is strongest below 3 Hz but spread as far as 10 Hz, highlighting the strong vibrato and clear articulation in this piece that carry the slower rhythmic structure.

Unlike conventional methods for computing the modulation spectrum (traditionally confined to a transform in the temporal dimension, discussed in Sect. 12.2.3), the two-dimensional modulation spectrum highlights *both* the spectral and temporal dynamics of the signal as well as the time alignment of these modulation patterns (i.e., cross-channel modulation phase), which is an important component for understanding spoken material and music compositions (Greenberg and Arai 2001; Hepworth-Sawyer and Hodgson 2016). The combined profile—across time and frequency—is the only mapping able to highlight subtle patterns in the original envelopes, such as frequency-modulations (FM), which are key signatures of many patterns in transitional speech sounds (e.g., diphthongs, semi-vowels), and metallic or percussive bell sounds (Chowning 1973).

Because of its span of the joint time-frequency space, the spectrotemporal modulation power spectrum (MPS) representation has been used as a dashboard to explore the precise loci of modulation energy driving the perception of speech and music. Recent work examined detailed tiling of the spectrotemporal modulation spectrum using granular techniques that inspected the perceptual contribution of various regions or building-blocks of the two-dimensional modulation profile. These methods, originally developed in vision research, aim to assign a quantifiable contribution of specific modulation energies to perceptual recognition of sound constructs using an approach referred to as “bubbles” (Gosselin and Schyns 2001). Because the spectrotemporal modulation profile is in a fact an image with temporal modulations on the *x*-axis and spectral modulations on the *y*-axis, the adoption of vision techniques can be seamlessly applied. These approaches have shown that the intelligibility of speech signals depends significantly on both spectrotemporal modulations that carry considerable modulation energy in the signal as well as those that carry linguistically relevant information (Venezia et al. 2016). A similar observation has also been reported for musical instrument recognition where low spectral and temporal modulation are the most salient regions to correlate with musical timbre, though signatures of individual instruments can be clearly discerned in the MPS space (Thoret et al. 2016). Alternative tiling techniques that use filtering (low-pass, notch filters) as well as dimensionality reduction and scaling have also been used to explore the informative regions of the MPS space (Elliott and Theunissen 2009; Elliott et al. 2013).

Overall, the MPS representation is proving to be a powerful descriptor of sound identity and timbre representation. It is also a space where joint interactions across time and frequency can be readily discerned. Still, it is not a very intuitive mapping of the acoustic waveform because it is a representation derived from the signal via at least two (typically more) transformations: from the acoustic signal to a time-frequency spectrogram and then to a time-frequency modulation spectrum (in addition to computing magnitude, power, binning operations, etc.). The models employed to perform these transformations do shape the salient details of the modulation profile and can invariably emphasize different aspects in this mapping, be it stimulus energies or perceptual energies.

The representation shown in Fig. 12.1 employs a straightforward two-dimensional Fourier transform to the time-frequency spectrogram. Other approaches have been proposed, including the use of two-dimensional wavelet transforms (Anden and Mallat 2014), bio-mimetic affine transforms mimicking receptive fields in mammalian auditory cortex (Chi et al. 2005), or even physiologically recorded receptive fields from single neurons in primary auditory cortex (Patil et al. 2012). Naturally, incorporating nonlinearities as reported in auditory processing can further color the readout of such modulation profiles, though limited work has been done that can shed light on the biological and perceptual relevance of nonlinearly warping the modulation space (as discussed in Sec. 12.5).

One of the other limitations of the modulation spectrum stems from the fundamental limit in precision by which modulations can be measured simultaneously in time and frequency. Much like the uncertainty principle is applied in a time-frequency spectrogram, the same is true in the modulation domain, which is effectively a transformation of the original space. The uncertainty principle, or Heisenberg principle, articulates the trade-off that one can achieve when attempting to represent time and frequency with infinite precision (Cohen 1995; Grochenig 2001). The smaller the window in time used to perform the analysis, the larger the bandwidth of spectral resolution afforded by this analysis because these two quantities have effectively a fixed product. Similarly, the temporal and spectral modulations derived within these constraints are also restricted relative to each other and, as such, provide a limited view of the spectrotemporal modulation profile of a signal (Singh and Theunissen 2003). How the brain deals with these limitations remains unknown, though they may explain the multi-resolution mapping of modulation in auditory cortical networks, as discussed in Sec. 12.3.

12.2.3 The Temporal Modulation Spectrum

As discussed in Sect. 12.2.2, the notion of modulation aims at identifying the patterns of inflection or change imposed on a signal. While the formal definition of such change does not necessarily identify the dimension on which it needs to operate, there is a large body of work that has focused on the temporal envelope. The temporal envelope is the main carrier of rhythmic fluctuations in the signal, and therefore its timescale and timespan are crucial information-bearing components of the signal. It is important to note that modulations along frequency also play a crucial role (as mentioned in Sect. 12.2.2; this issue will be expanded further in Sect. 12.4). Still, the temporal profile has garnered particular interest because of its simple mathematical derivation yet powerful importance in speech and music perception (Patel 2008).

The temporal modulation spectrum is obtained through a series of transformations that pass a signal *x[n]* through a bank of *M* bandpass filters in order to derive the envelope of each filter output. While this process is traditionally done on band-limited signals at the output of each filter, the premise of the computation does not preclude using broadband signals nor does it confine the bandwidth of the filterbank to a specific range. Naturally, the fluctuations of the filter outputs will be dictated by the choice of filterbank parameters, bandwidths, and frequency span.

Techniques used in the literature vary from using simple Fourier-like spectral decompositions (e.g., Fig. 12.1) to more perceptually grounded spectral mappings based on critical bands or a Bark scale (Moore 2003). The output of this filterbank analysis is an array of M filter outputs:

$$x_m\left[n\right]; m = 1, \ldots, M$$

The fluctuations of these output signals are then further isolated using an envelope extraction technique (either using the Hilbert transform or other transformations such as half-wave rectification and low-pass filtering), which results in a smooth envelope of each filter output ($E[x_m]$) whose variations are bounded both by the original bandwidth of the filterbank as well as the constraints of the envelope-tracking technique (Lyons 2011). Typically, this process is followed by a nonlinear mapping that compresses the linear envelope output using a nonlinear scaling function, such as square, logarithm, or a biologically motivated nonlinearity-mimicking nonuniform gain compression in the activation of the auditory nerve (Yang et al. 1992; Zhang et al. 2001). The compression is also used to counter the strong exponential nature of envelope amplitudes in natural sounds (Attias and Schreiner 1997).

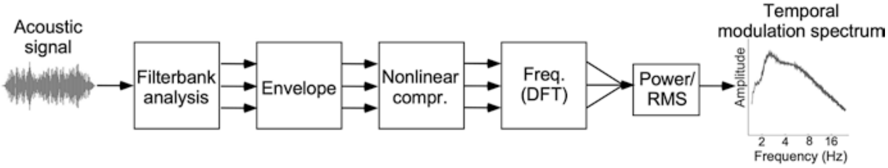

Fig. 12.2 Schematic of processing stages to derive the temporal modulation spectrum from an acoustic signal. The acoustic signal undergoes an initial analysis to map it onto a time-frequency representation before transformations of this spectrogram extract a temporal modulation spectrum from the envelope across different frequency channels. *DFT*, discrete Fourier transform; *RMS*, root-mean-squared

The readout of the fluctuations in the envelope signal is then obtained in the Fourier domain by mapping the time-domain signals onto a frequency-axis profile that is then summed across channels and transformed into power, root-mean-squared energy, or compressed magnitudes (Fig. 12.2).

Historically, this approach has been developed in the room acoustics literature via the concept of a modulation transfer function (MTF) (Houtgast and Steeneken 1985) and thus has relied on modulation filters employed to analyze the energy in the envelope signal at specific modulation points chosen along a logarithmic scale. An equivalent readout can be obtained using linearly spaced filters or by directly employing a Fourier transform on the compressed envelope signals. In either case, the resulting profile can then be combined across frequency bands and properly binned and scaled to yield an amplitude modulation spectrum that reflects envelope energies along different modulation frequencies. A major underlying assumption in this transformation is that such modulation frequencies of interest are below the pitch range, focusing primarily on the true envelope patterns or slow fluctuations in the signal. A number of constraints in the design of the processing steps must be considered in order to avoid artifacts or distortions that could mislead the readout of the spectrum profile (Qin Li and Les Atlas 2005).

12.2.4 Domain-Centric Representations

Some approaches have considered more structured analyses of the signal. In the case of speech sounds, the source-filter model of speech production has led to widely used techniques such as *Linear Predictive Coding* (LPC) (Schroeder and Atal 1985; see Caetano, Saitis, and Siedenburg, Chap. 11). The approach builds on the minimal but powerful simplification of speech production as a coupling of a vibrating source that generates the carrier signal with a filter that colors this carrier, hence giving speech its spectral shapes. As such, being able to decompose the signal into these two fundamental components disentangles the voicing characteristics primarily present in the source from the timbral cues primarily shaped by the filter, though there is a strong interaction between the two. From a linear systems point of view, separating the source (glottal signal) from the system (parameters of the vocal tract) means that the current speech sample can be closely approximated as a linear combination of past samples (hence the name linear predictive coding) (Rabiner and Schafer 2010). While an oversimplification of the complex dynamics of speech production, LPC modeling offers an indirect, yet effective, account of the spectral modulations shaping phonetic tokens of speech signals, though the temporal dynamics are often ignored by assuming the system (vocal tract) is quasi-stationary over short periods of time that span the analysis window.

A similar decomposition of source and filter cues underlies the widely popular cepstral decomposition, which provides a transformation of the filter characteristics in the *cepstral domain*. The cepstrum (a rotated version of the word spectrum) is an application of homomorphic signal processing techniques that apply a nonlinear mapping to a new domain wherein two components of a signal can be disentangled or deconvolved (Rabiner and Schafer 2010). Applied to speech signals, the *power cepstrum of a signal* is defined as the squared magnitude of the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of a signal (Caetano, Saitis, and Siedenburg, Chap. 11). Effectively, the cepstrum domain separates the slowly varying envelope (or modulation) signal from the rapidly varying excitation carrier signal, allowing the analysis of each component separately. The result is cepstral coefficients (and the related mel-frequency cepstral coefficients or MFCC) that offer an effective account of phoneme-dependent signal characteristics (Childers et al. 1977). Much like LPC, cepstral analysis remains limited to static representation of short segments of the speech signal (typically of the order of a phoneme) and focuses solely on the spectral characteristics of the modulating envelope.

Other approaches have been used to extend these representations to the time domain by computing derivative and acceleration over time, often referred to as delta and delta-delta coefficients of the signal, in an attempt to capture some of the temporal dynamics in the speech signal driven by prosodic and syllabic rhythms. While derivatives are rather simplistic extensions to capture the intricate temporal structure of the vocal tract during speech production, techniques such as LPC and MFCC remain powerful tools that provide a basic bread-and-butter analysis of speech signals with a formidable impact on many applications of speech analysis (Chen and Jokinen 2010; Hintz 2016). Their popularity speaks to the tremendous redundancies in speech signals as well as the powerful impact of a simple source-filter model in capturing some of the nuances of how speech signals are shaped and how they carry information.

While this source-filter view is rather unique to the speech production system, it is also applicable and quite popular for music analysis (Collins 2009; also see Caetano, Saitis, and Siedenburg, Chap. 11). Many musical instruments can be viewed as pairings of a source (a vibrating object such as a string) coupled with a filter (the body of the instrument that shapes the sound produced). Unlike a unitary model of source-filter analysis in speech, a common production system cannot be applied across instruments since the production may depend on vibrating strings, membranes, or air columns. As such, the distinction between the source and the filter is not as distinct as it is in speech and poses some challenges when applied to music signals, especially for non-Western music or polyphonic music (Muller et al. 2011).

While many approaches for music analysis borrow from a long tradition of speech processing, a number of elegant techniques have been developed specifically for music analysis particularly applied to domains of pitch, harmony, beat, tempo, and rhythm. The modulatory fluctuations in music, of both the spectral profile as well as the temporal envelope, have inspired a number of clever decompositions of music in order to hone in on the modulatory fluctuations in the signal. Some of these techniques extend the concept of a temporal modulation spectrum across multiple time scales. For instance, a family of modulation spectra spanning fast tempi (called *meter vectors*) offer a hierarchy of modulation spectra that summarizes the temporal patterning of events in a music signal nested across multiple time constants (Schuller 2013).

Overall, the analysis of modulations in speech and music signals is often informed by particular aspects of signal perception or production under study or with the ultimate goal of identification, recognition, or tracking. As such, the field enjoys a wide variety of tools developed from different perspectives that represent various facets of modulation. Ultimately, the modulation spectrum (in its many forms) has rather direct neurophysiological interpretations, as discussed in Sec. 12.3, though the elucidation of the exact substrate of specific forms of modulation encoding remains an open area of research.

12.3 Neurophysiological Interpretation of Modulations

The mapping of the informative acoustic attributes of an incoming signal takes different forms and varying levels of granularity as the signal is analyzed along the auditory pathway (Eggermont 2001). As early as cochlear processing, a sound signal entering the ear is decomposed along many bandpass frequency regions that span the basilar membrane, resulting in a time-frequency representation much like a short-term Fourier transform. The intricate details of sensory hair cell transduction shape the response across cochlear channels through a number of processing stages often modeled using half-wave rectification, low-pass filtering, and nonlinear compressions (Yang et al. 1992; Ibrahim and Bruce 2010). This process, analogous to deriving the envelope of an analytic signal using the Hilbert transform, effectively tracks the temporal variations of the signal along different frequency bands, which not only highlights the overall temporal patterns of the signal but specifically underscores the profiles of onsets and sustained activity as well as rhythmic changes (e.g., temporal cross-sections in Fig. 12.1).

The details in this temporal profile are encoded with gradually lower resolutions along the auditory pathway where the neural code appears to be increasingly selective to the slower dynamics that modulate the signal profile (Miller et al. 2002;

Escabi and Read 2003). This selectivity is reflected in the tuning parameters of neurons from the midbrain all the way to primary auditory cortex. Neural tuning characteristics are typically summarized using spectrotemporal receptive fields (STRF) (Elhilali et al. 2013). The STRF is a powerful tool in studying the selectivity of neurons to particular patterns in the stimulus. It typically treats a neuron as a system with a known input (the sound stimulus) and a measured output (the neural response). As is common in systems theory, the characteristics of a system (i.e., the system function) can be derived from its input and output or a class of inputs and corresponding outputs. This system function allows one to think of a neuron as a filter with a STRF that reflects the characteristics of the stimulus that best induces a strong response.

This STRF representation has been invaluable in shedding light on tuning characteristics of neurons along the auditory pathway. Of particular interest to the current discussion is the selectivity of neurons at the level of auditory cortex. While there is a great deal of variability across species and cortical layers, most auditory cortical neurons are sensitive to slow temporal and spectral modulation patterns (Depireux et al. 2001; Liu et al. 2003) commensurate with scales and dynamics of interest in modulation profiles, as discussed in Sect. 12.2. Unlike tuning in peripheral auditory nuclei, which captures mostly tonotopic energy across frequency, cortical neurons exhibit tuning sensitivity across at least three dimensions: (1) best frequencies (BF) that span the entire auditory range; (2) bandwidths that span a wide range from very broad (~2 octaves) to narrowly tuned (< 25% of an octave) (Schreiner and Sutter 1992; Versnel et al. 1995); and (3) temporal modulation dynamics that range from very slow to fast (1–30 Hz) (Miller et al. 2002).

Interpreting this representation from the vantage point of signal modulations, neural responses of a whole population of cortical neurons are mostly driven by temporal dynamics in the signal that are commensurate with the sound envelope (< 30 Hz). As a population, ensemble tuning of cortical neurons can therefore be tied to the temporal modulation spectrum of natural and complex sounds (Depireux et al. 2001; Miller et al. 2002). Complementing this axis are the spectral dynamics of the neural response across a cortical ensemble of neurons, which also spans spectral energies typical in signals with a characteristic resonance structure (extended over many octaves), that are able to extract harmonic and subharmonic structures in the spectrum (Schreiner and Calhoun 1995; Kowalski et al. 1996). The spectral selectivity of cortical neurons appears to match rather well the distinctive profile of spectral shapes in natural sounds, supporting the theory of a faithful alignment between acoustic modulation energy and neural encoding of such spectral modulations, which ultimately guides processing and perception of complex sounds (Leaver and Rauschecker 2010). Taking both dimensions into account, the considerable match between the modulation spectrum (derived directly from a signal corpus) and the tuning characteristics of an ensemble of cortical STRFs has been argued in the literature as possible evidence for the underlying role of the mammalian auditory cortex in encoding information-bearing components of complex sounds (Singh and Theunissen 2003).

While this view—at the ensemble level—reveals a formidable match between the acoustic properties of the signal and cortical neural tuning, the details of how these contours are derived are important to bear in mind because they impose a number of constraints on the modulation profiles under study and their interpretations. As mentioned earlier, the STRF is commonly interpreted through a systems theory view that deduces a system function based on the mapping between the input stimulus and the recorded neural response. Given interest in a system function that spans both time and frequency, a spectrotemporal representation of the stimulus is often preferred. However, the exact signal processing transformation used to map the spectrotemporal space dictates, to a great degree, the view and details emerging about the STRF. For instance, the tiling of the time-frequency space, the detailed resolution or sampling of such space, and the scaling of the amplitude energy profile of the stimulus can greatly affect the readout emerging from the neural mapping of this transformation and its match to the brain responses induced by complex acoustic stimuli.

Of particular interest is whether the use of a wavelet-based representation (based on logarithmic filter spacing) versus a spectrogram approach (akin to a Fourier transform) is more informative about the modulation spectrum and its neural underpinnings. On the one hand, wavelet-based analyses are generally preferred in explaining a number of perceptual findings, including modulation-tuning thresholds (Chi et al. 1999), given the closer biological realism in mimicking the frequency resolution provided by the auditory periphery. On the other hand, the time-frequency resolution tradeoff allows more modulation dynamics at the higher frequency bands of a wavelet representation and could magnify the effect of faster temporal dynamics. As such, linearly spaced filters have been preferred for deriving modulation spectra, especially when considering the temporal dynamics (Jepsen et al. 2008; Elliott and Theunissen 2009).

Though it is difficult to objectively quantify and compare the adequacy of different time-frequency mappings, a common technique used in the literature is to assess the goodness-of-fit for different mappings. A report by Gill et al. (2006) performed a systematic study of sound representations in an effort to elucidate the importance of certain factors in the derivation of neuronal STRFs. The study examined a number of parameters, particularly the use of linear versus logarithmic spacing of modulation filters, in deriving the time-frequency representation of the signal. Gill et al. (2006) found little evidence for a clear advantage in using linear versus logarithmic filter tiling for the derivation of time-frequency spectrograms of the stimulus and, subsequently, for the goodness-of-fit models of auditory neurons in the songbird midbrain and forebrain.

In contrast to the different ways of spectral tiling, which show little to no effect, Gill et al. (2006) reported stronger effects of adaptive gain control and amplitude compression of the stimulus in assessing auditory tuning. Those two aspects reflect the need for nonlinear transformations (both static and dynamic) in characterizing the neural underpinnings of auditory tuning to sound modulations. Nonlinear mappings of the time-frequency profile of the stimulus not only reflect the complex nature of neural processing along the auditory pathway, they also highlight the multiplexed layers of information-bearing components of natural sounds (Santoro et al. 2014). Reducing the concept of modulations to an envelope riding on top of a carrier is too simple to explain its role in timbre perception, especially for complex sounds.

12.4 How Informative are Modulations?

12.4.1 Modulations in Speech

What does the speech modulation spectrum reveal about understanding spoken language? Work dating a few decades back showed that comprehension of speech material is highly impaired in acoustic environments where distortions attenuate energies between 2–8 Hz (Steeneken and Houtgast 1979; Houtgast and Steeneken 1985). Those observations were further corroborated by later work in different languages that showed a dramatic decline in intelligibility if the integrity of the temporal modulation profile of speech was altered (with operations such as low-pass or bandpass filtering) (Drullman et al. 1994; Arai et al. 1999). Similar distortions disrupting the integrity of the spectral modulation profile by phase jitter or bandpass filtering are also equally detrimental to intelligibility, even if they do not alter the temporal envelope profile of speech (Arai and Greenberg 1998; Elhilali et al. 2003). In contrast, numerous studies have argued that any manipulations of speech that do not disrupt the integrity of its spectrotemporal modulations are harmless to its intelligibility (Shannon et al. 1995; Zeng et al. 2005). All in all, there is growing evidence that the spectrotemporal features captured by the speech MPS (see Sect. 12.2.2) offer a representation that closely maintains the phonetic identity of the sound as perceived by human listeners (Elliott and Theunissen 2009). The fidelity of the speech MPS correlates closely with intelligibility levels of speech in the presence of ambient noise and other distortions (Elhilali and Shamma 2008). The more a noise distorts the speech MPS, the more the decline of speech intelligibility. Conversely, noises that fall outside the core acoustic energy of the speech MPS have little effect on its intelligibility levels (Carlin et al. 2012).

The role of the spectrotemporal modulations of speech as information-bearing components has been leveraged extensively to sample speech signals for many applications, particularly automatic speech recognition (ASR) in the presence of background noise. Modulation-based analysis has enjoyed a lot of success as front-ends for ASR systems. Most studies have focused on the temporal evolution of the signal envelope to quantify modulation spectra (Kingsbury et al. 1998; Moritz et al. 2011), or estimations of the envelope *pattern* using temporal envelopes (Hermansky and Sharma 1999; Morgan et al. 2004), or using frequency-domain linear prediction (FDLP) (Athineos and Ellis 2003; Ganapathy et al. 2010). Also, a few attempts have been made to extend the analysis of modulations to both spectral and temporal domains; these studies have focused mainly on using two-dimensional Gabor filters (or other variants) as localized features for analysis of speech (Kleinschmidt 2003; Meyer et al. 2011). Across all of these different representations, the common thread is that once the speech signal is mapped onto a space that directly highlights its modulation content, the fidelity of that representation is sufficient to maintain the speech content and facilitate its robust recognition (Nemala et al. 2013). As such, this robustness provides empirical corroboration that such envelope modulations are indeed important information-bearing components of speech.

A faithful representation of speech signals has direct relevance for hearing prosthetics, particularly cochlear implants (CI), for which the fidelity of the signal has direct perceptual implications for the user (for more on timbre perception by CI users, see Marozeau and Lamping, Chap. 10). Speech modulations along the spectral axis are of particular interest in the case of cochlear implants because they dictate the resolution of the frequency axis and, ultimately, the channel capacity of the prosthetic device. Numerous studies have reported minimal disruption of speech comprehension in noise-free environments when only a few frequency channels are present over a range of hundreds of Hertz below 4 kHz (Shannon et al. 1995). Importantly, as few as four channels (i.e., a spectral resolution as low as 1.6 cycles/octave) are sufficient to maintain intelligibility. Such resolution is generally too low for acceptable levels of speech recognition in noise and also results in impoverished music perception (as discussed in Sec. 12.4.2). By the same token, it has been argued that fully resolving formant spectral peaks (up to 2 cycles/octave) results in great improvement in intelligibility, especially when speech is corrupted with noise (Friesen et al. 2001; Elliott and Theunissen 2009). The tradeoff between the spectral resolution sufficient for speech perception in quiet settings and the spectral resolution necessary for speech recognition in the presence of noise remains a matter of debate (Friesen et al. 2001; Croghan et al. 2017). This is especially important given the variability across listeners in their ability to utilize the spectrotemporal cues available to them.

The debate over modulations and spectrotemporal resolutions necessary for speech perception highlight the fact that there is more to speech than just its envelope (Moore 2014). While the view of modulations as an envelope fluctuation riding a fast carrier is true to a great extent, that view conceals the complex role played by the underlying fast structure of the signal in complementing the representation, and ultimately the perception, of speech signals. The temporal fine-structure and spectral details play key roles in speech perception in noise (Qin and Oxenham 2003; Shamma and Lorenzi 2013), sound localization (Smith et al. 2002), lexical-tone perception (Xu and Pfingst 2003), repetition or residue pitch perception (deBoer 1976), and fundamental frequency discrimination (Houtsma and Smurzynski 1990). Psychophysical evidence suggests that one of the advantages that normal subjects have over hearing-impaired listeners is improved local target-to-masker ratios, especially in the presence of spectrally and temporally fluctuating backgrounds (Peters et al. 1998; Qin and Oxenham 2003). The notion of listening in the spectral and temporal "dips" of the masker sounds is less realizable for hearing impaired listeners because of poor spectral selectivity and reduced temporal resolution (Glasberg and Moore 1992).

Fine details of speech (especially along the spectrum) are also crucial for dealing with stationary and narrowband noises and pitch-centric speech processing

(McAuley et al. 2005; Wang and Quatieri 2012). Hence, one has to be careful in interpreting the perceptual salience of the slow envelope for speech perception as an exhaustive account of the speech signal. Reducing the speech signal to a dichotomy consisting of two independent components—envelope and fine-structure—is a flawed premise. The envelope and fine-structure components are not only impossible to tease apart, but they also convey complementary information about the speech signal, especially in everyday listening environments (Shamma and Lorenzi 2013).

12.4.2 Modulations in Music

Much like speech, music signals carry a multiplexed and highly layered structure of dynamics both spectrally and temporally. Music perception evokes a complex experience that spans multiple elements that include pitch, melody, timbre, and rhythm among others. The representations of signal modulations in their different forms directly encode many facets of these musical attributes (see Caetano, Saitis, and Siedenburg, Chap. 11). Among musical elements, modulations have a very tight affiliation with the perception of timbre both in terms of sound identity but also as musical quality.

Acoustically, a musical note is shaped by the physical constraints of the instruments as well as the motor control of the player. These constraints whittle the acoustic signal with modulatory envelopes that carry some of the timbral properties of music. The acoustic signature of these constraints naturally shapes both spectral and temporal profiles of the acoustic signal, and they ultimately inform the perceptual experience as these cues are decoded by the auditory system. Numerous perceptual studies have shed light on these acoustic correlates (McAdams, Chap. 2; Agus, Suied, and Pressnitzer, Chap. 3) with spectrum as the most obvious candidate. The spectral shape of a musical note is naturally shaped by the vibration mode and resonances of the instrument and that modulates not only the spectral energy profile but also frequency peaks, spectral sharpness and brightness, amplitudes of harmonic partials, spectral centroid, and spectral irregularities. The temporal envelope of the signal is also heavily modulated, and correlates of timbre can also be gleaned from the energy buildup, onset information, attack over time, and the spectral flux over time. All these attributes, spanning both spectral and temporal modulations, not only determine the identity of a musical instrument but also the perceived timbral quality of musical-instrument sounds.

In a study directly relating spectrotemporal modulations to the perception of timbre, Patil et al. (2012) explored the fidelity of neural activation patterns in mammalian auditory cortex in accurately replicating both classification of musical instruments as well as perceptual judgements of timbre similarities. The study examined the ability of a cortical mapping to reflect instrument-specific characteristics. Patil et al. (2012) specifically assessed whether a processing pipeline that mimicked the transformation along the auditory pathway up to primary auditory cortex was able to capture the instrument identity from a wide variety of isolated notes from eleven instruments playing 30–90 different pitches with 3–10 playing styles, 3 dynamic levels, and 3 manufacturers for each instrument (an average of 1980 tones per instrument). The model was able to distinguish the identity of different instruments with an accuracy of 98.7%, corroborating the hypothesis that timbre percepts can be effectively explained by the joint spectrotemporal analysis performed at the level of mammalian auditory cortex.

Patil et al. (2012) also examined a more stringent constraint to explore how well this cortical mapping reflected distances between instruments that correlated with the perceptual judgements of timbre similarity by human listeners. In other words, it is not sufficient to judge whether a timbre representation is able to distinguish a violin from a cello, but can it also discern that a violin is perceived as more similar to a cello than it is to a bassoon. The representation based on spectrotemporal receptive fields was indeed able to project notes from individual instruments onto a space that maintains their relative distances according to similarity judgements of human listeners. The faithful representations of spectrotemporal modulations in the cortical space were correlated with human similarity judgements with an accuracy of r = 0.944.

While the relation between spectrotemporal modulation tuning at the level of primary auditory cortex and timbre perception is quite strong, it is important to note a number of observations. The fact that the timbre space spans a complex interplay of spectral and temporal dimensions is not surprising and has been established through a large body of work spanning many decades (see Siedenburg, Saitis, and McAdams, Chap. 1). What timbre analysis via a biomimetic cortical model sheds light on is the fact that the decoding of acoustic modulations along both time and frequency over a rich representational space appears to be necessary and sufficient to almost fully capture the complete set of acoustic features pertinent to instrument identity and timbre similarity. It also pushes forth the debate about the cardinality of a timbre space, one that extends beyond few descriptors to require a high number of dimensions. This direct relationship between modulations and timbre perception reinforces the theories tying modulation with information-bearing components of the musical signal.

One of caveats to this theory (that the study by Patil and colleagues brought to light) is that the modulation space cannot be a separable one, spanning marginally along time and frequency (Patil et al. 2012). Rather, the *joint* representation along both directions is crucial, emphasizing spectrotemporal dynamics in the timbre profile (see McAdams, Chap. 2). For instance, frequency modulations (FM), such as vibrato, impose rich dynamics in music signals, and they can only be discerned reliably by examining the joint spectrotemporal space. The role of spectrotemporal modulations that underly music perception has been directly reported using psychophysical studies that correlate music perception abilities and modulation detection thresholds for time alone, frequency alone, and joint time-frequency (Choi et al. 2018). The correlations are stronger with spectrotemporal modulation-detection thresholds, further corroborating the idea that the configuration of the timbre space directly invokes a modulation space based on *joint* spectrotemporal dynamics (Elliott et al. 2013).

Another important observation from the Patil et al. (2012) study is that timbre representation in a biomimetic spectrotemporal modulation space is only effective at replicating human judgements when augmented by a nonlinear mapping boundary. A number of studies, in fact, have established this nonlinear behavior, especially at the level of auditory cortex, as it pertains to encoding of complex sound patterns (Sadagopan and Wang 2009). The exact nature, neural underpinnings, and the specificity of this nonlinearity to different sound classes remain unclear. As such, the quest for a direct mapping between spectrotemporal modulations and a timbre space remains unfulfilled.

12.4.3 Common and Unique Modulation Profiles in Speech and Music

As one examines the relationship between modulation profiles and the perception of speech and music, a natural question that arises pertains to commonalties and differences between profiles of these two sound classes. While temporal dynamics of speech are widely diverse and multiscale (e.g., variations across speakers, languages, prosodic profiles), variations in musical temporal patterns are even more diverse across genres, performances, and arrangements (Patel 2008). An analysis of modulation temporal profiles contrasting speech with Western musical samples shows drastic differences between these two sound classes (Fig. 12.3). This analysis, reproduced from (Ding et al. 2017), depicts temporal modulation profiles obtained by computing a discrete Fourier transform (DFT) of narrowband power envelope signals representing the root-mean-squared of the outputs of cochlear channels that correspond to four frequency bands. This processing contrasts the dynamics of a speech corpus, consisting of nine languages (American English, British English, Chinese, Danish, Dutch, French, German, Norwegian, and Swedish), against datasets of Western music samples that include classical music by single-voice string instruments and multi-voice instruments, symphonic ensembles, jazz, and rock (for details, see Ding et al. 2017). Immediately notable is the shift in the peak temporal modulation between speech and music. While speech has the now established peak around 4–8 Hz (typically attributed to physical dynamics of speech production articulators), the music dataset analyzed in this study shows visibly lower peaks with a plateau between 0.5–3 Hz. A number of physical and perceptual constraints can offer some explanations for the disparity. The kinematics of hand movements in music production (for the Western samples analyzed) impose a natural constraint on the temporal rates of movement with a preferred frequency of arm movements at around 1.5 Hz (Van Der Wel et al. 2009). There is also a relationship between emergent temporal modulations of music signals and underlying beats of the musical phrasing that also tend to highlight a rate of 1.5–3 Hz (van Noorden and Moelants 1999).

In addition to temporal modulation profiles, the distinction between speech and musical sounds is also very prominent with respect to their spectral profiles. Speech perception remains effective even over rather coarse sampling of the spectral axis. A case in point is the effectiveness of cochlear implants at conveying intelligible

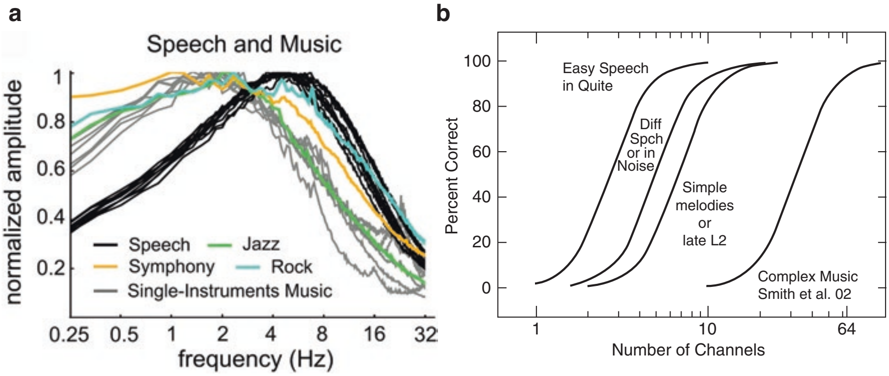

Fig. 12.3 Modulation profiles in speech and music. (**A**) The modulation spectrum of speech (*black*), single-instrument (*gray*), and multi-part music (*colors*). (**B**) Meta-analysis incorporating results across many studies to examine speech and music recognition (*y-axis*) as a function of the number of spectral channels (*x-axis*) in a noise band vocoder (**A** reprinted from Ding et al. 2017; **B** reprinted from Shannon 2005; both used with permission from Elsevier)

speech with very few channels, at least in favorable listening conditions (Wilson 2004). That is far from being the case for music perception (McDermott 2004), for which poor spectral resolution directly impacts melody recognition as well as timbre perception, two crucial aspects of the complex experience that constitutes music perception. Fig. 12.3 reproduces an illustration by Shannon (2005) that highlights the effects of spectral resolution on the perception of speech and music signals in addition to the effect of difficulty of listening. Panel B provides a meta-analysis across a number of studies that examine speech and music recognition rates as a function of the number of spectral channels in a noise-band vocoder. Speech detection in quiet listening conditions is contrasted with the same task under more challenging situations (including more difficult sentences, background noise, recognition in a second language, etc.). The trends show a clear need for improved spectral resolution under challenging conditions. This requirement for finer spectral resolution is further underscored when a task of melody recognition in the presence of competing melodies is used. This latter study results in the interesting contrast between speech versus melody identification: as low as 3 channels to achieve 75% correct identification of speech sentences in quiet listening conditions to as high as 40 channels to achieve 75% correct identification of melodies (Smith et al. 2002).

An interesting question regarding the distinction between spectral and temporal modulations of speech and music signals is how the perceptual system integrates across these modulation cues. For speech signals, joint spectrotemporal modulations capture temporal fluctuations of certain spectral peaks (e.g., formant transitions or speech glides). But work on automatic speech recognition suggests that joint spectrotemporal modulations are not necessary to improve recognition of words in the presence of distortions (Schädler and Kollmeier 2015).

These results argue that capturing signal transitions along both time and frequency may be less crucial for recognizing speech in noise. Instead, a reduced representation of spectral and temporal modulations (separately) is argued to yield comparable recognition as the joint-modulation representation. Unfortunately, there have been limited extensions of this exploration to definitely rule out a role of joint spectrotemporal modulations in speech recognition.

In contrast, the role of joint spectrotemporal modulations in musical timbre has been clearly demonstrated. There is strong evidence that a separable space, spanning time and frequency separately, is insufficient to capture the nuances of timbre required for distinguishing the timbre of different musical instruments. Instead, a modulation representation of both time and frequency axes is important to explicitly encode key musical constructs such as frequency modulations common in string vibrato (Patil et al. 2012; Elliott et al. 2013).

The divergence in acoustic attributes of both sound classes offers a potential rationale for different neural circuits that underlie the processing of speech and music in the brain (Zatorre et al. 2002; Norman-Haignere et al. 2015). The left hemisphere plays a more prominent role in complex linguistic functions; whereas, the right hemisphere appears to notably favor tasks involving tonal patterns or spectral processing, two aspects that are most related to the perception of music (Liégeois-Chauvel et al. 1998). This specialization beyond auditory cortex builds on an underlying common circuitry of mid-level and primary cortical representations that appear to focus primarily on extracting spectrotemporal modulations in incoming complex sound patterns. These very modulations appear to be a crucial backbone needed to carry information about complex sounds such as speech and music.

12.5 Summary

Theoretically, modulation is nothing but a mapping of an acoustic signal that highlights its fluctuations or indicates how its energy changes over time and frequency. These modulations are shaped by the source from which the signal emanates; hence, they can inform about the physics of that source and ultimately the signal's timbre. In practice, however, quantifying modulations is a nontrivial endeavor that takes many formulations and interpretations. Modulations of complex signals, such as speech and music, are a multifaceted construct that varies along multiple time scales and granularities, and they are shaped as much by the physics of the source as by the neural representations of acoustic energy in the brain. This chapter reviews some of the common representations of modulations and reflects on their perceptual and neural interpretation.

A number of questions surrounding the representation and role of modulations remain open. For example, what is the contribution of nonlinearities, which are pervasive in brain networks, in shaping the encoding of signal modulations in the auditory system? As discussed throughout this chapter, most constructs of modulations rely on transformation of the signal energy to another domain via spectral mappings such as the Fourier transform. These transformations maintain operations in the original vector space of time-frequency and, as such, are limited in their ability to manipulate or warp the mapping of spectrotemporal modulations. This is also true in the case of biomimetic constructs, such as spectrotemporal receptive fields, used to analyze neural activity in the central auditory system (Depireux and Elhilali 2013). While the receptive field view of auditory processing offers a rich set of tools to explore the encoding of sound characteristics, they are very much limited by approximative assumptions of linearity that are often compensated for in backend systems by means of nonlinear kernels that are often used in machine learning (Hemery and Aucouturier 2015). Understanding these nonlinearities is not only essential in the study and modeling of brain networks but also crucial to truly grasp the role played by sound modulations in informing perception.

The encoding of modulations is likely to be further shaped by active engagement in listening tasks and deployment of cognitive processes, notably attention. These top-down processes are known to greatly modulate neural encoding of incoming signals (Shamma and Fritz 2014), yet their role in shaping the representation of signal modulations remains largely unknown. Future research efforts addressing these questions will shed light on aspects of modulations that the brain hones in on when listening in multisource environments, for instance, their function in helping the auditory system deal with the cocktail party problem (Elhilali 2017).

Acknowledgements Dr. Elhilali's work is supported by grants from the National Science Foundation (NSF), the National Institutes of Health (NIH), and the Office of Naval Research (ONR).

Compliance with Ethics Requirements Mounya Elhilali declares she has no conflicts of interest.

References

Anden J, Mallat S (2014) Deep scattering spectrum. IEEE Trans Signal Process 62:4114–4128. https://doi.org/10.1109/TSP.2014.2326991

Arai T, Greenberg S (1998) Speech intelligibility in the presence of cross-channel spectral asynchrony. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, p 933–939

Arai T, Pavel M, Hermansky H, Avendano C (1999) Syllable intelligibility for temporally filtered LPC cepstral trajectories. J Acoust Soc Am 105:2783–2791

Athineos M, Ellis DPW (2003) Frequency-domain linear prediction for temporal features. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU):261–266

Attias H, Schreiner CE (1997) Temporal low-order statistics of natural sounds. In: Adv. Neural Inf. Proc. sys. (NIPS). MIT Press: Cambridge, MA, p 27–33

Carlin MA, Patil K, Nemala SK, Elhilali M (2012) Robust phoneme recognition based on biomimetic speech contours. In: Proceedings of the 13th annual conference of the international speech communication association (INTERSPEECH), p 1348–1351

Chen F, Jokinen K (2010) Speech technology: theory and applications, 1st edn. Springer, New York

Chi T, Gao Y, Guyton MC, Ru P, Shamma S (1999) Spectro-temporal modulation transfer functions and speech intelligibility. J Acoust Soc Am 106:2719–2732

Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am 118:887–906

Childers DG, Skinner DP, Kemerait RC (1977) The cepstrum: a guide to processing. Proc IEEE 65:1428–1443. https://doi.org/10.1109/PROC.1977.10747

Choi JE, Won JH, Kim CH, Cho Y-S, Hong SH, Moon IJ (2018) Relationship between spectrotemporal modulation detection and music perception in normal-hearing, hearing-impaired, and cochlear implant listeners. Sci Rep. 8(1). https://doi.org/10.1038/s41598-017-17350-w

Chowning JM (1973) The synthesis of complex audio spectra by means of frequency modulation. J Audio Eng Soc 21:1–10

Cohen L (1995) Time-frequency signal analysis, 1st edn. Prentice-Hall, Englewood Cliffs

Collins N (2009) Introduction to computer music, 1st edn. Wiley, Chichester/West Sussex

Croghan NBH, Duran SI, Smith ZM (2017) Re-examining the relationship between number of cochlear implant channels and maximal speech intelligibility. J Acoust Soc Am 142:EL537–EL543. https://doi.org/10.1121/1.5016044

deBoer E (1976) On the "residue" and auditory pitch perception. In: Keidel W, Neff D (eds) Auditory system (handbook of sensory physiology). Springer, Berlin, pp 479–583

Depireux DA, Elhilali M (eds) (2013) Handbook of modern techniques in auditory cortex. First. Nova Science Publishers, Inc., New York

Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85:1220–1234

Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D (2017) Temporal modulations in speech and music. Neurosci Biobehav Rev 81:181–187

Divenyi P, Greenberg S, Meyer G (eds) (2006) Dynamics of speech production and perception. IOS Press, Amsterdam, p 388

Drullman R, Festen JM, Plomp R (1994) Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 95:1053–1064

Dudley H (1939) Remaking speech. J Acoust Soc Am 11:169–177

Dudley H (1940) The carrier nature of speech. Bell Syst TechJ 19:495–513

Eggermont JJ (2001) Between sound and perception: reviewing the search for a neural code. Hear Res 157:1–42

Elhilali M (2017) Modeling the cocktail party problem. In: Middlebrooks J, Simon JZ, Popper AN, Fay RR (eds) The auditory system at the cocktail party. Springer, New York, pp 111–135

Elhilali M, Shamma S (2008) Information-bearing components of speech intelligibility under babble-noise and bandlimiting distortions. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 4205–4208

Elhilali M, Chi T, Shamma SA (2003) A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Commun 41:331–348. https://doi.org/10.1016/S0167-6393(02)00134-6

Elhilali M, Shamma SA, Simon JZ, Fritz JB (2013) A linear systems view to the concept of STRF. In: Depireux D, Elhilali M (eds) Handbook of modern techniques in auditory cortex. Nova Science Pub Inc, New York, pp 33–60

Elliott TM, Theunissen FE (2009) The modulation transfer function for speech intelligibility. PLoS Comput Biol 5:e1000302

Elliott TM, Hamilton LS, Theunissen FE (2013) Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. J Acoust Soc Am 133(1):389–404. https://doi.org/10.1121/1.4770244

Escabi MA, Read HL (2003) Representation of spectrotemporal sound information in the ascending auditory pathway. Biol Cybern 89:350–362

Freeman R (2004) Telecommunication system engineering, fourth edn. Wiley-Interscience, New York

Friesen LM, Shannon RV, Baskent D, Wang X (2001) Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J Acoust Soc Am 110:1150–1163

Ganapathy S, Thomas S, Hermansky H (2010) Robust spectro-temporal features based on autoregressive models of Hilbert envelopes. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 4286–4289

Gill P, Zhang J, Woolley S, Fremouw T, Theunissen F (2006) Sound representation methods for spectro-temporal receptive field estimation. J Comput Neurosci 21:5. https://doi.org/10.1007/s10827-006-7059-4

Glasberg BR, Moore BC (1992) Effects of envelope fluctuations on gap detection. Hear Res 64:81–92

Gosselin F, Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. Vis Res 41(17):2261–2271. https://doi.org/10.1016/S0042-6989(01)00097-9

Greenberg S (2004) Temporal properties of spoken language. In: Proceedings of the international congress on acoustics. Kyoto, Japan, p 441–445

Greenberg S, Arai T (2001) The relation between speech intelligibility and the complex modulation spectrum. In: Proceedings of the 7th European conference on speech communication and technology (Eurospeech-2001), p 473–476

Grochenig K (2001) Foundations of time-frequency analysis. Birkhauser, Boston

Hemery E, Aucouturier J-J (2015) One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis. Front Comput Neurosci 9(80). https://doi.org/10.3389/fncom.2015.00080

Hepworth-Sawyer R, Hodgson J (2016) Mixing music, First edn. Routledge, New York/London

Hermansky H, Sharma S (1999) Temporal patterns (TRAPs) in ASR of noisy speech. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 292

Hintz M (2016) Digital speech technology:pProcessing, recognition and synthesis. Willford Press

Houtgast T, Steeneken HJM (1985) A review of MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J Acoust Soc Am 77:1069–1077

Houtsma AJM, Smurzynski J (1990) Pitch identification and discrimination for complex tones with many harmonics. J Acoust Soc Am 87:304–310

Ibrahim R, Bruce I (2010) Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues. In: Lopez-Poveda EA, Palmer AR, MR (eds) The neurophysiological bases of auditory perception. Springer, New York, pp 429–438

Jepsen ML, Ewert SD, Dau T (2008) A computational model of human auditory signal processing and perception. J Acoust Soc Am 124:422–438

Katz M (2006) The violin: a research and information guide. Routledge Taylor and Francis Group, London/New York

Kingsbury B, Morgan N, Greenberg S (1998) Robust speech recognition using the modulation spectrogram. Speech Commun 25:117–132

Kleinschmidt M (2003) Localized spectro-temporal features for automatic speech recognition. In: Proceedings of Eurospeech, p 2573–2576

Kowalski N, Depireux DA, Shamma SA (1996) Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. J Neurophysiol 76:3503–3523

Leaver AM, Rauschecker JP (2010) Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci 30:7604–7612

Qin Li, Les Atlas (2005) Properties for modulation spectral filtering. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), p 521–524

Liégeois-Chauvel C, Peretz I, Babaï M, Laguitton V, Chauvel P (1998) Contribution of different cortical areas in the temporal lobes to music processing. Brain 121:1853–1867. https://doi.org/10.1093/brain/121.10.1853

Liu RC, Miller KD, Merzenich MM, Schreiner CE (2003) Acoustic variability and distinguishability among mouse ultrasound vocalizations. J Acoust Soc Am 114:3412–3422

Lyons RG (2011) Understanding digital signal processing, third edn. Prentice Hall, Upper Saddle River

McAuley J, Ming J, Stewart D, Hanna P (2005) Subband correlation and robust speech recognition. IEEE Trans Speech Audio Process 13:956–963. https://doi.org/10.1109/TSA.2005.851952

McDermott HJ (2004) Music perception with cochlear implants: a review. Trends Amplif 8:49–82

Meredith D (ed) (2016) Computational music analysis. Springer International Publishing, Cham

Meyer B, Ravuri S, Schaedler M, Morgan N (2011) Comparing different flavors of spectro-temporal features for ASR. In: Proceedings of the 12th annual conference of the international speech communication association (INTERSPEECH), p 1269–1272

Miller LM, Escabí MA, Read HL, Schreiner CE, Escabi MA, Read HL, Schreiner CE (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87:516–527. https://doi.org/10.1152/jn.00395.2001

Moore BCJ (2003) An introduction to the psychology of hearing, 5th edn. Emerald Group Publishing Ltd, Leiden

Moore BCJ (2014) Auditory processing of temporal fine structure: Effects of age and hearing loss, 1st edn. World Scientific Publishing, Co, Hackensack/New Jersey

Morgan N, Chen BY, Zhu Q, Stolcke A (2004) Trapping conversational speech: extending TRAP/tandem approaches to conversational telephone speech recognition. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 40 vol.1

Moritz N, Anemuller J, Kollmeier B (2011) Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), p 5492–5495

Müller M (2015) Fundamentals of music processing. Springer International Publishing, Cham

Muller M, Ellis DPW, Klapuri A, Richard G (2011) Signal processing for music analysis. J IEEE, Sel Top Signal Process 5:1088–1110. https://doi.org/10.1109/JSTSP.2011.2112333

Nemala SK, Patil K, Elhilali M (2013) A multistream feature framework based on bandpass modulation filtering for robust speech recognition. IEEE Trans Audio Speech Lang Process 21:416–426. https://doi.org/10.1109/TASL.2012.2219526

Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88:1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Patel AD (2008) Music, language, and the brain, First edn. Oxford University Press, Oxford

Patil K, Pressnitzer D, Shamma S, Elhilali M (2012) Music in our ears: the biological bases of musical timbre perception. PLoS Comput Biol 8:e1002759. https://doi.org/10.1371/journal.pcbi.1002759

Peters RW, Moore BC, Baer T (1998) Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. J Acoust Soc Am 103:577–587

Pickett JM (1999) The acoustics of speech communication: fundamentals, speech perception theory, and technology. Allyn & Bacon, Boston

Poeppel D, Idsardi WJ, van Wassenhove V (2008) Speech perception at the interface of neurobiology and linguistics. PhilosTransR Socl B BiolSci 363:1071–1086

Qin MK, Oxenham AJ (2003) Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J Acoust Soc Am 114:446–454

Rabiner L, Schafer R (2010) Theory and applications of digital speech processing, First edn. Pearson, Upper Saddle River

Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. Philos Trans R Soc B-Biological Sci 336:367–373

Sadagopan S, Wang X (2009) Nonlinear spectrotemporal interactions underlying selectivity for complex sounds in auditory cortex. J Neurosci 29:11192–11202

Sadie S (ed) (2001) The new grove dictionary of music and musicians, Second edn. Macmillan, London

Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, Formisano E (2014) Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput Biol 10(1). https://doi.org/10.1371/journal.pcbi.1003412

Schädler MR, Kollmeier B (2015) Separable spectro-temporal Gabor filter bank features: reducing the complexity of robust features for automatic speech recognition. J Acoust Soc Am 137:2047–2059. https://doi.org/10.1121/1.4916618

Schreiner C, Calhoun B (1995) Spectral envelope coding in cat primary auditory cortex: properties of ripple transfer functions. J Audit Neurosci 1:39–61

Schreiner CE, Sutter ML (1992) Topography of excitatory bandwidth in cat primary auditory cortex: single-neuron versus multiple-neuron recordings. J Neurophysiol 68:1487–1502

Schroeder M, Atal B (1985) Code-excited linear prediction(CELP): high-quality speech at very low bit rates. In: Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), p 937–940. doi: https://doi.org/10.1109/ICASSP.1985.1168147

Schuller B (2013) Applications in intelligent music analysis. Springer, Berlin/ Heidelberg

Shamma S, Fritz J (2014) Adaptive auditory computations. Curr Opin Neurobiol 25:164–168. https://doi.org/10.1016/j.conb.2014.01.011

Shamma S, Lorenzi C (2013) On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. J Acoust Soc Am 133:2818–2833. https://doi.org/10.1121/1.4795783

Shannon RV (2005) Speech and music have different requirements for spectral resolution. Int Rev Neurobiol 70:121–134

Shannon R, Zeng F, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304

Singh N, Theunissen F (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am 106:3394–3411

Smith ZM, Delgutte B, Oxenham AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. Nature 416:87–90. https://doi.org/10.1038/416087a

Steeneken HJ, Houtgast T (1979) A physical method for measuring speech-transmission quality. J Acoust Soc Am 67:318–326

Thoret E, Depalle P, McAdams S (2016) Perceptually salient spectrotemporal modulations for recognition of sustained musical instruments. J Acoust Soc Am 140(6). https://doi.org/10.1121/1.4971204

Turner RE, Sahani M (2011) Demodulation as probabilistic inference. IEEE Trans Audio Speech Lang Process 19(8):2398–2411

Van Der Wel RPRD, Sternad D, Rosenbaum DA (2009) Moving the arm at different rates: slow movements are avoided. J Mot Behav 42:29–36. https://doi.org/10.1080/00222890903267116

van Noorden L, Moelants D (1999) Resonance in the perception of musical pulse. J New Music Res 28:43–66. https://doi.org/10.1076/jnmr.28.1.43.3122

Venezia JH, Hickok G, Richards VM (2016) Auditory "bubbles": efficient classification of the spectrotemporal modulations essential for speech intelligibility. J Acoust Soc Am 140(2):1072–1088. https://doi.org/10.1121/1.4960544

Versnel H, Kowalski N, Shamma SA (1995) Ripple analysis in ferret primary auditory cortex. III. Topographic distribution of ripple response parameters. J Audit Neurosci 1:271–286

Wang TT, Quatieri TF (2012) Two-dimensional speech-signal modeling. IEEE Trans Audio Speech Lang Process 20:1843–1856. https://doi.org/10.1109/TASL.2012.2188795

Wilson BS (2004) Engineering design of cochlear implants. 20:14–52

Xu L, Pfingst BE (2003) Relative importance of temporal envelope and fine structure in lexical-tone perception. J Acoust Soc Am 114:3024–3027

Yang X, Wang K, Shamma SA (1992) Auditory representations of acoustic signals. IEEE Trans Inf Theory 38:824–839

Zatorre RJ, Belin P, Penhune VB (2002) Structure and function of auditory cortex: music and speech. Trends Cogn Sci 6:37–46

Zeng F-G, Nie K, Stickney GS, Kong Y-Y, Vongphoe M, Bhargave A, Wei C, Cao K (2005) Speech recognition with amplitude and frequency modulations. Proc Natl Acad Sci 102:2293–2298

Zhang X, Heinz MG, Bruce IC, Carney LH (2001) A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. J Acoust Soc Am 109:648–670

Chapter 13
Timbre from Sound Synthesis and High-Level Control Perspectives

Sølvi Ystad, Mitsuko Aramaki, and Richard Kronland-Martinet

Abstract Exploring the many surprising facets of timbre through sound manipulations has been a common practice among composers and instrument makers. The digital era radically changed the approach to sounds thanks to the unlimited possibilities offered by computers, which made it possible to investigate sounds without physical constraints. In this chapter, we describe investigations on timbre based on the analysis-by-synthesis approach, which consists of using digital synthesis algorithms to reproduce sounds and further modify the parameters of the algorithms to investigate their perceptual relevance. In the first part of the chapter, timbre is investigated in a musical context. An examination of the sound quality of different wood species used to make xylophones is first presented. Then the influence of physical control on instrumental timbre is described in the case of clarinet and cello performances. In the second part of the chapter, investigations of environmental sounds are presented to identify invariant sound structures that can be considered as the backbone or the bare minimum of the information contained in a sound that enables the listener to recognize its source both in terms of structure (e.g. size, material) and action (e.g. hitting, scraping). Such invariants are generally composed of combinations of audio descriptors (e.g., decay, attack, spectral density, and pitch). Various investigations on perceived sound properties responsible for the evocations of sound sources are identified and described through both basic and applied studies.

Keywords Action-object paradigm · Analysis by synthesis · Digital interfaces
Intuitive synthesis control · Musical interpretation · Sound semiotics

S. Ystad (✉) · M. Aramaki · R. Kronland-Martinet
CNRS, Aix Marseille University, PRISM (Perception, Representations, Image, Sound, Music), Marseille, France
e-mail: ystad@prism.cnrs.fr; aramaki@prism.cnrs.fr; kronland@prism.cnrs.fr

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_13

13.1 Introduction

13.1.1 Historical Overview of Timbre and Instrumental Control

Timbre has been one of the main concerns for instrument makers, musicians, and composers throughout history. Certain instruments, such as organs, were particularly well adapted to exploring various timbres due to the numerous pipes that made it possible to combine different harmonics. In the eighteenth century, Dom Bedos de Celles, a French Benedictine monk and organ maker, published a treatise entitled *The Art of the Organ-Builder* (*L'art du facteur d'orgues*) in which he not only describes the principles behind organ building but also the many ways of imitating certain instrumental timbres by adding or removing sounds of pipes tuned to multiples of the fundamental frequency (Dom Bedos 1766) (for a detailed description of organ structure and timbre, see Angster et al. 2017). Such techniques, which constitute a practical application of the Fourier (1878) theorem on periodic functions, are claimed to have been used as early as in the fifteenth century, that is, four centuries before Fourier published his fundamental theorem showing that sounds can be reconstituted by a sum of harmonics and before Helmholtz (1868) related timbre to the proportion of harmonic amplitudes.

When electricity became viable for use in technology thanks to Faraday in 1831, inventors started to build new musical instruments often based on the *additive synthesis technique*, which consists of creating sounds by adding elementary signals, typically sine functions with different frequencies and amplitudes. Phonic wheels, with which the harmonics of the sounds could be added and removed to imitate timbres of both known and unknown instruments, were used to develop the Telharmonium in 1897. The Hammond organ developed in the 1930s was based on the same principle but with new control features. With the B-3 model, which offered control of the attack time (Caetano, Saitis, and Siedenburg, Chap. 11), the instrument suddenly became extremely attractive to jazz musicians due to its new means of adjusting the degree of percussiveness of the sounds (De Wilde 2016). Another invention that focused on the possibilities of controlling timbre variations was the "ondes Martenot," which was based on high-frequency (radio) waves (similarly to the more widespread theremin). This instrument was equipped with a six-octave keyboard, a sliding metal ring that enabled the performer to produce glissandi, as well as a drawer with timbre controls that made it possible to switch between different waveforms (e.g., sinusoidal, triangle and square waves, pulse waves, and noises) and to route the instrument's output to various loudspeakers providing either reverberation effects, sympathetic resonances, or "halo" effects (creation of diffuse sound fields).

Yet another instrument that offered a huge palette of new timbres was the modular Moog synthesizer developed in the 1960s, which enabled the creation of sounds using four basic modules, namely oscillators, amplifiers, filters, and envelopes. By offering fine envelope control of the attack, release, sustain, and decay parts of the sound, an extremely rich and subtle timbre control was made possible for the musician. Unfortunately, the many control possibilities were not intuitive and made the first versions of the instrument difficult to use. These examples illustrate musicians' and composers' passionate quest for new timbres, which was nicely expressed by the composer Edgar Varèse (1917, p. 1): "I dream of instruments obedient to my thought and which with their contribution to a whole new world of unsuspected sounds, will lend themselves to the exigencies of my inner rhythm" (translated from the French by Louise Varèse).

13.1.2 Timbre Studies Induced by the Digital Era

In spite of the many amazing instruments dedicated to analog synthesis (obtained from electric pulses of varying amplitude), the arrival of the digital era, in which the computer was introduced, revolutionized our conception of musical sounds and perception. In 1957, Max Mathews developed the first sound synthesis computer program (MUSIC I) at the Bell Labs in the USA, which he used to create the first computer-generated musical piece in history (Mathews 1963). The use of sound synthesis enables one to generate an infinite number of sounds without being constrained by physics. Several pioneers in the field, such as Jean-Claude Risset, David Wessel, and John Chowning (all both composers and scientists), rapidly seized the opportunity to use this new tool as a means to establish links between perception and sound structures by developing an analysis-by-synthesis approach in which the reconstruction of the sound became the criterion for the relevance of the analysis. It was by such an approach that Risset (1965), for example, revealed the importance of the temporal evolution of different spectral components in trumpet sounds; his study pointed out that the increase in spectral bandwidth as a function of amplitude is linked to the brassy effect of the instrument. Similarly, Mathews et al. (1965) managed to improve the realism of the attack of bowed string instruments by introducing frequency variations to synthetic sounds.

The analysis-by-synthesis approach also was used in the first studies on the perceptual representation of timbre proposed by Grey (1977). The study involved constructing synthetic emulations of musical instruments in which certain parts of the signal were degraded through simple transformations. Through listening tests, certain acoustic parameters (e.g., the attack time and spectral centroid) were identified as relevant from a perceptual point of view. More recent studies based on either resynthesized sounds obtained from the analysis of recorded sounds or from synthesized sounds that are not necessarily perfect imitations of the original sound have revealed several audio descriptors (Peeters et al. 2011) that are representative of specific sound categories (McAdams, Chap. 2; Saitis and Weinzierl, Chap. 5).

At this stage one might think that Varèse's dream of instruments that give access to any timbre that a composer imagines would be available due to the many studies that have established links between sound categories and audio descriptors. This is true in theory but not that easy in practice, since our ability to describe and control sound structures to obtain given timbres is limited. In fact, digital sound synthesis is based on low-level parameters such as amplitudes and frequencies of spectral components and their temporal evolution. This signal content is a consequence of the physical behavior of the source and does not necessarily reflect how the sound is perceived (McAdams, Chap. 2). A major challenge in the domain of digital synthesis, therefore, is to unveil the sound structures that are responsible for the recognition of the sound source (e.g., size, shape) and the sound-producing action (e.g., hitting, scraping) in order to be able to reproduce and control such evocations in an intuitive manner. This means that various scientific disciplines must be associated in order to link perceptual and cognitive experiments with physical characteristics of sounds.

13.1.3 Source Recognition and Timbre

Several composers, psychologists, musicians, and scientists have worked on human perceptions of environmental sounds. During the late 1940s, the French scientist, philosopher, and musician Pierre Schaeffer introduced a new musical genre that he called "musique concrète" in the "Studio d'essai" of the French public radio (RTF). This new trend consisted of distributing recorded sounds for which the source could not be easily recognized over loudspeakers in order to favor reduced or *acousmatic listening*, thereby forcing listeners to focus on the sound itself and not on the source that created the sound. Schaeffer realized that the specification of the physical structure of the sound was not adequate to control the auditory effects because "all music is made to be heard" (Schaeffer 2017, p. 97), and the relation between the physical signal and the perception of musical sounds at the time was grossly insufficient from his viewpoint (also see Schaeffer 1966).

Schaeffer's ideas can be found in later studies. For instance, Smalley (1994) introduced the term *source bonding* as "the natural tendency to relate sounds to supposed sources and causes and to relate sounds to each other because they appear to have shared or associated origins." Gaver (1993) distinguished what he called "ecological or everyday listening" (hearing events per se) from analytical or musical listening (focusing on intrinsic sound properties as in the case of Schaeffer's reduced listening). Gaver also took his inspiration from Gibson (1979), who introduced the ecological approach to perception in the visual domain. This theory supposes that our perception is direct, without any influence of inference or memory, and is based on the recognition of specific signal morphologies, which can be considered as invariant structures that transmit the perceptual information. In addition to Gaver (1993), several other authors have adapted (or at least partially integrated) the ecological approach for use in the auditory domain (Warren and Verbrugge 1984; McAdams and Bigand 1993). The notion of invariant sound structures is particularly interesting for sound synthesis and control purposes since the identification of such structures makes it possible to focus on evocative sound structures to produce sounds and sound metaphors, and it enables intuitive or high-level control of sounds from semantic descriptors (e.g., small, big, metal, wood, hollow, plain).

Accordingly, this chapter is a discussion of timbre from the point of view of sound synthesis and control. In the first part, timbre and musical sounds are investigated in various cases. First, wood species are evaluated by a xylophone maker in terms of sound quality. Then the link between instrumental control and timbre is examined in the case of clarinet and cello performances. In the second part, environmental sounds are explored through studies based on brain imaging techniques centered on semiotics of sounds, that is, how meaning is attributed to sounds. Then the implications of audio descriptors are examined with regard to the identification of invariant sound structures responsible for the evocation of sound sources and events. Finally, particular mapping strategies between low-level synthesis parameters, audio descriptors, and semantic labels describing the sound sources for intuitive high-level control of sounds are considered.

13.2 Timbre Studies Based on Analysis-Synthesis Approaches in Musical Contexts

This section deals with timbre-related questions about musical instruments, in particular, the quality of musical sounds and the role of timbre in instrumental control and musical performances.

13.2.1 Timbre-Based Wood Selection by a Xylophone Maker

The mechanical properties of wood species strongly influence the sound quality. When choosing their wood species, xylophone makers carefully listen to the sounds they produce. Little is known about the relationship between the sound quality and the physical parameters characterizing wood species or the criteria used to choose wood. Aramaki et al. (2007) studied the perceptual criteria. For this purpose, a professional xylophone maker was asked to evaluate samples of different tropical and subtropical wood species with the same geometry. Sounds produced by these samples were first recorded and classified by the instrument maker through a free classification test. Then the sounds were resynthesized and tuned to the same pitch before the same instrument maker performed a new classification. Statistical analyses of both classifications revealed the influence of pitch on the xylophone maker's judgements and pointed out the importance of two audio descriptors: *frequency-dependent damping* and *spectral bandwidth*, indicating that the instrument maker searched for highly resonant and crystal-clear sounds. These descriptors can be further related to physical and anatomical characteristics of wood species, thereby providing recommendations for choosing attractive wood species for percussive instruments. Previous studies relating auditory cues to geometry and material properties of vibrating objects have pointed out the importance of internal friction related to the damping factors of the spectral components (Lutfi and Oh 1997; Giordano and McAdams 2006) as hypothesized by Wildes and Richards (1988). Other studies on the sound quality of musical instruments have been performed for violins (Saitis et al. 2012, 2017). These studies were based on quantitative analyses of violinists' preference judgements during playing tasks and psycholinguistic analyses of their spontaneous verbalizations describing the playing experience, and they led to a model linking auditory and haptic sensations to the timbre, quality, and playability of the instrument (cf. Saitis et al. 2018).

13.2.2 Timbre Control in Clarinet Performances

Investigations on musical timbre do not solely focus on the mechanical properties of the instrument itself but also on the way a musician can control timbre while playing the instrument. The following section describes a study on the influence of a clarinet player's pressure and aperture on the resulting timbre, using a physical synthesis model (Barthet et al. 2010a), which is followed by an investigation of the influence of timbre on expressiveness in clarinet performance (Barthet et al. 2010b).

To draw a link between the control parameters and the resulting timbre in clarinet performance, a synthesis model was used to generate perfectly calibrated sounds (Guillemain et al. 2005). Fifteen sounds obtained by different values of reed aperture and blowing pressure were evaluated through dissimilarity ratings. The statistical analyses of the perceptual evaluations resulted in a timbre space with dimensions that correlated well with attack time and spectral centroid for the first dimension, the energy ratio between odd and even harmonics for the second dimension, and the energy of the second to fourth harmonics for the third dimension (second tristimulus coefficient). A correlation between the control parameters and the timbre space could also be found, revealing that the pressure control correlated well with the third dimension and had a strong influence on the odd/even ratio. Furthermore, the reed aperture was well correlated with the first dimension and with the attack time and the spectral centroid (see Fig. 13.1). These results allowed for the prediction of the instrumental timbre from the values of the control parameters. Hence, small values of reed opening and blowing pressure result in long attack times and low spectral centroid values, while increasing reed aperture induces increases in the odd/even ratio. For more information on the acoustics of wind instruments, refer to Moore (2016) and Wolfe (2018).

Studies on musical performance have revealed rhythmic and intensity deviations with respect to the musical score, leading to proposals of various musical rules (Sundberg 2000). Although timbre variations are likely to be used by musicians as a means to add expressivity to the performance, they have been more or less ignored, probably for two reasons: (1) scores do not contain timbre specifications; and (2) timbre variations strongly depend on the specificities of each instrument and, therefore, might be difficult to integrate with general performance rules. In the study by Barthet et al. (2010b), timbre variations were analyzed in

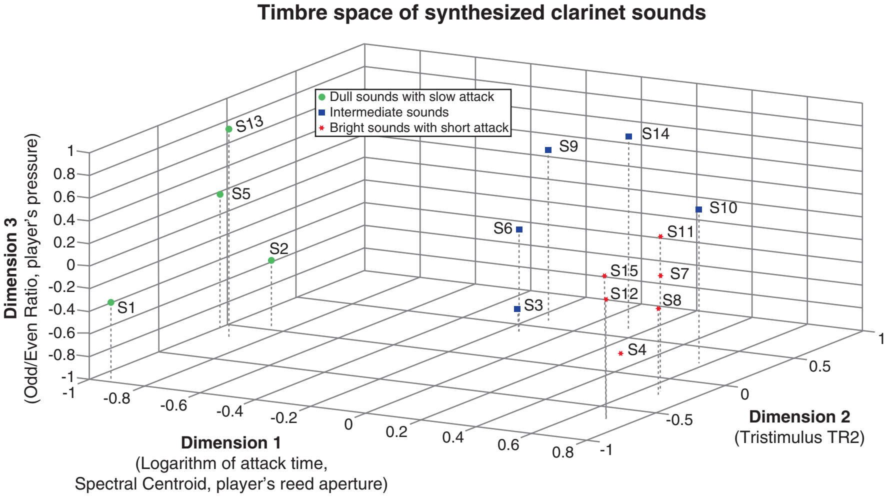

Fig. 13.1 Timbre space of clarinet sounds for different reed apertures and blowing pressures. The reed aperture is correlated with Dimension 1 (Logarithm of attack time, spectral centroid), whereas the pressure is correlated with Dimension 3 (Odd/Even Ratio)

order to investigate their influence on expressiveness in clarinet performance. Mechanical and expressive clarinet performances of excerpts from Bach and Mozart were recorded. An objective performance analysis was then conducted, focusing on the acoustic correlates of timbre. A strong interaction between the expressive intentions and the audio descriptors (attack time, spectral centroid, odd/even ratio) was observed for both musical excerpts. The timbre-related changes across expressive levels did not occur at every note but were specific to some notes or groups of notes in the musical phrases (such as the first note in a phrase or specific passages). The most salient changes were in the mean spectral centroid and odd/even ratio values and in the range of variation in the durations of the tones. These changes seemed to be made more frequently in the case of long notes (such as half and quarter notes), possibly because a performer needs a certain time to control the timbre while playing.

In a companion study, Barthet et al. (2011) examined the perceptual influence of certain acoustic timbre correlates (spectral centroid, SC), timing (intertone onset interval, IOI), and intensity (root-mean-squared envelope) on listeners' preferences between various renderings. An analysis-by-synthesis approach was used to transform previously recorded clarinet performances by reducing the expressive deviations from the SC, the IOI, and the root-mean-squared envelope (dynamics). Twenty skilled musicians were asked to select which version (recorded versus transformed) they preferred in a paired-comparison task. The results showed that the removal of the SC variations most significantly decreased the musical preference of the performances. That finding indicates the transformation altered the original timbre of the clarinet tones (the identity of the instrument) and drastically affected the time-evolving spectral shapes, causing the tones to be static and not lively. This result suggests that acoustic morphology, which strongly depends on the context (i.e., the previous and following notes, a fast or slow musical tempo), is important to convey expressiveness in music. More recent studies have analyzed the combination of timbre from different musical instruments and have examined how the timbres of certain instruments can blend together and whether musicians consciously control blend during performances (Lembke et al. 2017a, b). Those studies revealed that musicians adjusted their timbre, in particular the frequencies of the main formant or spectral centroid, depending on whether they had a role as leader or follower during the performance. Other studies have explored the more general role of timbre in orchestration and musical tension and proposed a typology of orchestral gestures based on large-scale timbral and textural changes (Goodchild et al. 2017; McAdams, Chap. 8).

13.2.3 Timbre and Ancillary Gestures

Another aspect that appears to be important for musical expressiveness is the musician's movements during the performance. Are the sound-producing gestures solely responsible for the sound quality and expressiveness or do ancillary gestures that are not directly involved in the sound production also play a role? Several studies on ancillary gestures have been performed in the case of the clarinet (Wanderley et al. 2005; Desmet et al. 2012), the piano (Jensenius 2007; Thompson and Luck 2012), the harp (Chadefaux et al. 2013), and the violin (Van Zijl and Luck 2013). In the case of clarinet performances, the body movements of the musician generated amplitude modulations of partials of the sounds, which were often perceived as beating effects. Such modulations are essentially due to changes in directivity of the instrument that follows the ancillary gestures of the musician. In the case of piano performance, a circular movement of the elbow enables a larger displacement of the hand (Jensenius 2007). This gesture depends on parts of the body that are not directly implied in the instrumental gesture (Thompson and Luck 2012). In a study on ancillary gestures in cello performances (Rozé et al. 2016, 2017), professional cellists were asked to play a score as expressively as possible in four postural conditions. The four conditions were a normal condition (N), a mentally constrained condition in which the cellists were asked to move as little as possible (static mental), and two physically constrained conditions in which the torso was attached to the back of the chair with a race harness (static chest) and, for the most constrained condition (Fig. 13.2), both the head and torso were immobilized with the addition of a neck collar (static chest, head).

A musical score was divided into six parts based on cello exercises with specific technical difficulties. The tempo (45 beats per minute) was given by a metronome before the beginning of each session and two bowing modes (detached and legato) were compared. Sounds and body movements were recorded. The analyses of the

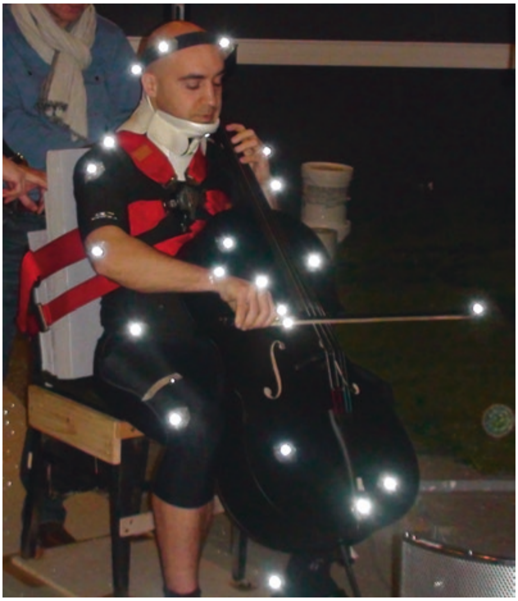

Fig. 13.2 Constrained postural condition in which the cellist is attached to the chair by a race harness, and his head is immobilized by a neck collar. The reflective markers on the picture enable body movement recordings by motion-capture cameras

performances revealed that, for certain notes, the timbre (and to a certain extent the rhythm) was modified in the fully constrained condition. In particular, in a specific passage of the score, a degradation of the timbre induced a noticeable perception of harshness. An analysis-by-synthesis approach associated with listening tests revealed that this phenomenon could be characterized by an energy transfer or a formant shift toward higher-order harmonics, a decrease in attack time, and an increase in fluctuations of harmonic amplitudes.

Based on those results, a predictive model of perceived harshness was proposed that depended on three audio descriptors: (1) the attack time; (2) the ratio between the first and second mel-frequency cepstral coefficients (MFCC), characterizing slow fluctuations of the spectral envelope; and (3) the harmonic spectral variation, reflecting the evolution of the energy of the harmonic components over time. The three-dimensional space resulting from this analysis presented tight analogies with the acoustic correlates of classical timbre spaces (McAdams, Chap. 2). The first dimension indicated that participants were particularly sensitive to spectral fluctuation properties (harmonic spectral variation), while the second and third dimensions, respectively, were well explained by spectral attributes (harmonic spectral centroid, the MFCC ratio) and a temporal attribute (the attack slope). This indicates that a greater brightness combined with a softer attack would contribute to an increase in the perceived harshness of a cello sound.

13.3 Semiotics of Environmental Sounds

In timbre research, musical sounds have been given a lot of attention since the first multidimensional representations proposed by Grey (McAdams, Chap. 2; Saitis and Weinzierl, Chap. 5). Fewer studies are available on environmental sounds, possibly because such sounds are hard to control and often complex to model from both physical and signal points of view. Although environmental sounds have been an important source of inspiration for composers throughout history, the focus in this section is not on the musical context but rather is on a pragmatic approach that considers the way meaning is perceived and attributed to environmental sounds.

In daily life, people are confronted with environmental sounds that are more or less consciously processed. Sounds tell us about the weather, living creatures in our surroundings, potential dangers, for example, which means that our environment constantly communicates information to us. How do people interpret and attribute meaning to such sounds? Can these sounds provide new ways to communicate if we manage to extract their perceptual essence and further implement it in sound synthesis processes? Can a common sense be attributed to such "environmental languages" that can be compared to the semantics of spoken languages?

As a first attempt to answer these questions, an investigation on the perception of isolated sounds would be interesting. One of the major issues that arises from the cognitive neuroscience point of view is whether similar neural networks are involved in the allocation of meaning in the case of language and that of sounds of other kinds. In a seminal study, Kutas and Hillyard (1980) showed that sentences that ended with words that were out of context (e.g., the fish is swimming in the river/carpet) elicited a larger negative amplitude of the evoked-related potential (ERP) component (measured on the scalp of the subjects 400 ms after the onset of the incongruous word: the N400 component) than when the last word was congruent. The N400 has been widely used since that time to study semantic processing in language. Authors of recent studies used a priming procedure with nonlinguistic stimuli such as pictures, odors, music, and environmental sounds (for reviews, see Aramaki et al. 2009; Schön et al. 2009). In the next section, two priming experiments are presented that used nonlinguistic stimuli to observe the negativity of ERP components for related versus unrelated stimuli. In the first case, priming effects induced by pairs of abstract sounds (favoring reduced listening) and by written words were investigated; in the second case, pairs of impact sounds evoking different material categories were examined.

13.3.1 Priming with Abstract Sounds

Although the results of previous priming experiments have been interpreted mostly as reflecting some kind of conceptual priming between words and nonlinguistic stimuli, they may also reflect linguistically mediated effects. For instance, watching a picture of a bird or listening to a birdsong might automatically activate the verbal label “bird”. The conceptual priming cannot be taken to be purely nonlinguistic, therefore, because of the implicit naming induced by processing the stimulus.

Certain studies have attempted to reduce the likelihood that a labeling process of this kind takes place. To this end, “abstract” sounds, which have the advantage of not being easily associated with an identifiable physical source, are useful (Schaeffer 1966; Merer et al. 2011). Sounds of this kind include environmental sounds that cannot be easily identified by listeners or that can give rise to many different interpretations that are dependent on the context. Abstract sounds also include synthesized sounds and laboratory-generated sounds in general, if their origin is not clearly detectable. Note that alarm or warning sounds do not qualify as abstract sounds, since they obey specific acoustic and emotional criteria (Bergman et al. 2009). In practice, making recordings with a microphone close to the sound source, using musical instruments in untraditional ways, or using everyday objects (e.g., tools or toys) are common ways of creating abstract sounds. Sound synthesis methods such as granular synthesis, which consists of adding many very short (typically 1–50 ms) sonic grains to form larger acoustic events (Roads 1988), are also efficient means of creating abstract sounds.

In the present study conceptual priming tests were conducted using word/sound pairs, for which the level of congruence between the prime and the target was varied. In the first experiment, a written word (the prime) was presented visually before an abstract sound (the target), and the participants had to decide whether or not the sound and the word matched. In the second experiment, the order of presentation was reversed. Results showed that participants were able to assess the relationship between the prime and the target in both sound/word and word/sound presentations, showing low intersubject variability and good consistency. The contextualization of the abstract sound facilitated by the presentation of a word reduced the variability of the interpretations and led to a consensus between participants in spite of the fact that the sound sources were not easily recognizable. Electrophysiological data showed the occurrence of an enhanced negativity in the 250–600 ms latency range in response to unrelated targets as compared to related targets in both experiments, suggesting that similar neural networks are involved in the allocation of meaning in the case of language and sounds. In addition, differences in scalp topography were observed between word and sound targets (from frontocentral to centroparietal distributions), which can be taken to argue that the N400 effect encompasses different processes and may be influenced by both the high-level cognitive processing of the conceptual relation between two stimuli and lower-level perceptual processes that are linked with the specific acoustic features of the sounds, such as attack time, spectral centroid, spectral variation, and others (Schön et al. 2009). This means that a combination of sound features, called invariants (cf. Sect. 13.4.2), might be used by listeners to determine specific aspects of sounds.

13.3.2 Priming with Material Categories

Pursuing this topic farther in a subsequent study, Aramaki et al. (2009) sought to completely avoid the use of words as primes or targets. Conceptual priming was therefore studied using a homogeneous class of nonlinguistic sounds (e.g., impact sounds) as both primes and targets. The degree of congruence between the prime and the target was varied in the following three experimental conditions: related, ambiguous, and unrelated. The priming effects induced in these conditions were then compared with those observed with linguistic sounds in the same group of participants.

Results showed that the error rate was highest with ambiguous targets, which also elicited larger N400-like components, than related targets in the case of both linguistic and nonlinguistic sounds. Aramaki et al. (2009) also found that N400-like components were activated in a sound-sound (rather than word-sound) design, showing that linguistic stimuli were not necessary for this component to be elicited. The N400-like component, therefore, may reflect a search for meaning that is not restricted to linguistic meaning. This study showed the existence of similar relationships in the congruity processing of both nonlinguistic and linguistic target sounds, thereby confirming that sounds can be considered as an interesting way to convey meaningful messages.

13.4 Toward Intuitive Controls of Sounds

The identification of perceptually relevant signal morphologies is of great interest in the domain of sound synthesis since it opens a new world of control possibilities. When computer music was at its very beginning in the 1960s, the famous scientist John Pierce made the following enthusiastic statement about sounds made from computers: “Wonderful things would come out of that box if only we knew how to *evoke* them” (Pierce 1965, p. 150, emphasis added). In spite of the many synthesis algorithms that have been developed based on signal models (Kronland-Martinet et al. 1997; Cook and Scavone 1999) or physical models (Bensa et al. 2003; Bilbao and Webb 2013) that provide a perfect resynthesis of sounds (no difference is perceived between the original and the synthesized sound), the issue of control is still a great challenge that prevents many potential users from considering sound synthesis in their applications. This issue has always interested composers and musicians, and a large number of interfaces and control strategies for digital sound synthesis have been proposed in the musical domain, starting with the pioneering works of Moog (1987) to more recent works (Cook 2001; Gobin et al. 2003). An overview of digital musical interfaces can be found in Miranda and Wanderley (2006).

The first works on perceptual control were presented by David Wessel (1979) who proposed a new way to navigate within a perceptual sound space based on timbre space (Grey 1977), defined by Krimphoff et al. (1994, p. 625) as “the mental organization of sound events at equal pitch, loudness and duration. The geometric distance between two timbres corresponds to their degree of perceived dissimilarity." Hence, instead of directly controlling the parameters of an additive synthesis algorithm, he defined a control of the audio descriptors that is correlated to the dimensions of the perceptual space, in particular the evolving spectral energy distribution and various temporal features (either the attack rate or the extent of synchronicity among the various components). Such a control was more intuitive than the control of basic signal parameters that defines synthesis algorithms, such as the frequencies and amplitudes of spectral components.

However, the sound manipulation remains difficult and insufficient for controls based on a direct description of the sound event (e.g. « the sound from a big wooden barrel rolling on gravels »). In this case a deeper analysis that enables to identify complex sound structures that are responsible for evoking sources and events is necessary to propose such intuitive controls based on verbal labels describing the perceived event. Such an approach necessitates a confrontation of distinct scientific domains: experimental psychology, cognitive neuroscience, acoustics, physics, and mathematics.

13.4.1 Evidence of Actions on Objects in Everyday Timbres

During the last 20 years, automobile manufacturers have shown an increasing interest regarding the influence of sounds on the perceived quality of cars. A large variety of sound sources have been investigated, such as the noises from flashing lights and alarms (Suied et al. 2008), the air conditioning system (Susini et al. 2004; Roussarie 2005), and car horns (Lemaitre et al. 2009). Such sounds carry a lot of information that a driver (or a passenger) uses unconsciously. In the following sections two studies are presented that were aimed at linking the signal properties of car-door sounds and motor noise to the listeners' impressions of automobile quality.

13.4.1.1 Door-Closure Sounds

The first study by Bezat (2007) was initiated by an automobile company whose employees noticed that the brief sound produced when slamming the car door was responsible for the customers' mental impressions of the quality and the solidity of the car—this sound even was important for car sales! It was quite surprising that such a brief sound (duration less than 250 ms) could have any influence on the customers. To understand how the sound influenced the customers, signal morphologies responsible for the evocation of solidity and quality of the car had to be found in order to propose a predictive model of this relationship (Bezat 2007). For this purpose, door-closure sounds obtained from recordings of car doors (different brands and car categories) were analyzed and evaluated. The perceptual judgements of car doors were obtained from different types of listening tests. Following Gaver's (1993) definition, both analytical and ecological listening were considered.

In the case of analytical (or musical) listening, the sound is described without reference to the event in order to reveal perceptually relevant signal morphologies useful for signal analysis. To incite participants to characterize sounds in this rather unnatural way, a sensory analysis method was used (Roussarie et al. 2004). For this purpose, naive subjects were trained during several sessions to define a minimal number of descriptors to qualify the sound they heard. Then the naïve subjects evaluated the stimuli. Through this method, a sensory profile could be obtained for each sound. The conditions for success of a sensory panel (consisting of the trained naïve subjects) are sound discriminability, consistency over time, and their consensus. This is the reason why such a long procedure is necessary to transform naive subjects into efficient judges in order to obtain their consensus on all the descriptors. This approach revealed that the door-closure sound is described mainly by the intensity (e.g., loud or soft) and by the onomatopoeias BONM (pronounced [bɔ̰̃m]) and KE (pronounced [kø]) as determined by the sensory panel. By comparing the analytical properties with expert listening, the BONM descriptor could be related to the low-frequency closure sound and the KE descriptor was related to the high-frequency contribution that characterizes the lock component in the car door signal.

In the case of ecological (or everyday) listening, the event associated with the sound characterized by a set of natural properties and the evoked associations were described. Listening tests with both naive and expert listeners were performed, revealing that naive listeners were able to discriminate the doors by their quality, solidity, energy of closure, door weight, and door closure effectiveness in a coherent manner. This is in line with previous studies (Kuwano et al. 2006) that resulted in coherent quality evaluations of car-door sounds across participants from semantic differential tests based on a predefined adjective scale (Saitis and Weinzierl, Chap. 5). The expert listeners identified the elements of the car door that contribute to the sound (the joints, the lock mechanism, and the door panel) in contrast to the more macroscopic evaluations of the naive listeners.

These different listening tests allowed the establishment of a network of perceptual properties of door-closure sounds (Fig. 13.3) that illustrated the links between the sensory panel's description of the sounds (i.e., its analytical properties), the notion of weight and closure linked to the natural properties and, finally, the evocations of quality and solidity of the car. The impressions of solidity and quality were linked to the sensation of a heavy door and a gentle gesture, which in turn was characterized by the members of the sensory panel as closing sounds without flaws (e.g., vibrations) that were low-pitched with little lock presence (strong BONM and weak KE) and of low intensity. In line with these results, the influence of the weight of the car door on the perceived quality was also confirmed by Scholl and Amman (1999) in a study in which car-door noises were evaluated after physical modifications of the car door sources.

An analysis-synthesis approach based on empirical mode decomposition (EMD) (Huang et al. 1998) was then applied to separate the perceived source contributions

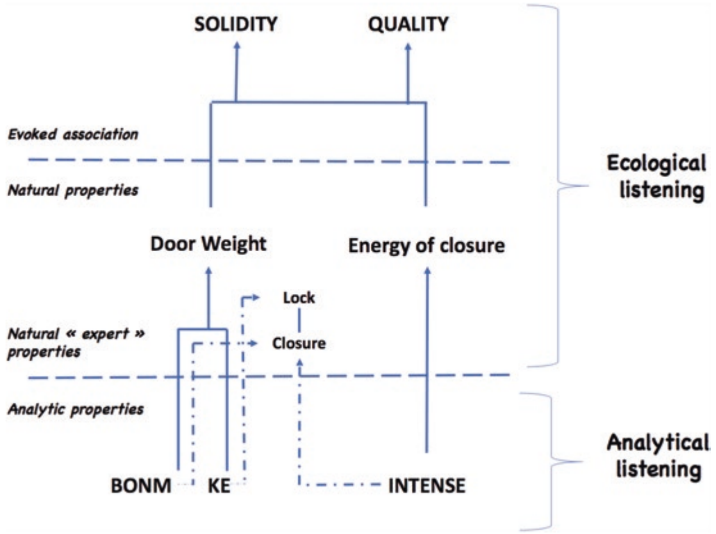

Fig. 13.3 Relation between the sensory descriptors (onomatopoeia terms: *BONM, KE*; plus *INTENSE*), the natural properties (*weight* and *energy closure*), and the evocations of *quality and solidity* of a car for sounds during door closure. The *bottom part* of the figure refers to the the way intrinsic sound properties are perceived through analytical listening. The *middle part* refers to the way the different sources are perceived through natural listening, while the *top part* refers to the global impression of the car. The evocations of a heavy door and a gentle gesture induced the sensation of a solid, high quality car

(lock and closure) contained in the sound. This method consists of identifying iteratively intrinsic mode functions (IMFs) of the signal (both amplitude and frequency modulated) and separating (on an oscillatory scale) fast components from slower ones. This is done by pointing out the located maxima and minima of the signal, constructing the superior and inferior envelopes, and then calculating the mean envelope. The first mode is thus obtained. The algorithm is then processed on the rest of the signal until the second mode is obtained. By finding EMD modes, the rest of the signal (the residue) has less and less extrema. The decomposition process stops when the last residue has only three extrema. The signal can then be reproduced perfectly by simple addition of the modes. Based on this signal analysis that enables separation of slow and rapid oscillations and the perceptual analyses, it could be concluded that an acceptable car-door noise should contain three impacts that evoke the latch mechanism (characterized as KE by the sensory panel) and one low-frequency impact that evokes the door impact (characterized as BOMN by the sensory panel).

An additive synthesis model based on exponentially damped sinusoids was then used to synthesize the sounds. By adjusting the amplitudes, the damping coefficients, and the time between the different impacts, car-door closure sounds corresponding to different vehicle qualities could then be generated. Further listening tests were run using the synthesized stimuli in order to relate the signal parameters to the perceived quality. The results indicated that the energy and the damping of the door impact (linked to the evocation of the weight of the door) and the elapsed time between the four impacts (related to the evocation of a well-closed door) were found to mediate the perception of solidity and quality of the car (Bezat et al. 2014). This study confirmed that signal invariants evoking the solidity and quality of a car can be identified.

13.4.1.2 Motor Sounds

Another challenging study by Sciabica (2011) that sought to relate evocations and sound structures was proposed by the same car company and concerned the perception of motor noise during acceleration. The aim was to characterize the dynamic behavior of motor noise timbre in terms of "sportiness" and further propose a perceptual control of a synthesis model related to the degree of sportiness.

Sounds perceived in car passenger compartments are the result of three acoustic sources: the engine sounds, the tire-road source, and the aerodynamic source. The sound from tire-road source is due to the interaction between the tires and the road and depends on three main parameters: car speed, tire texture, and road texture. The contact between tire and road generates low-frequency noise. The sound from the aerodynamic source is a broadband noise whose global sound level increases with speed. It mainly has a low-frequency energy distribution (below 400 Hz), but its perceptual contribution is also important in the high-frequency domain (up to 1000 Hz). Indeed, aerodynamic noise mainly masks high engine orders, but its impact can also be observed at low engine orders. The engine sound is complex with rich overtones: the fundamental frequency varies with the engine rotation speed, and the level of each harmonic depends on the multiple resonances inside the car. When the engine sound is sufficiently audible in the car, it can be described by perceptual attributes such as booming, brightness, and roughness. Booming is associated with a resonant low-frequency harmonic and can be considered as annoying for the driver (Chaunier et al. 2005). Increased brightness reflects the presence of audible high-order harmonics, whereas increased roughness reflects audible secondary harmonics that interact with the main harmonics. The resulting signal, therefore, is a mixture of several harmonics and a low-frequency broadband noise.

Although these perceptual attributes can be clearly identified at a given instant, they fail to properly characterize the dynamic variation of the car sounds during acceleration, for instance. Hence, the evocation of identity and perceived quality must be investigated in terms of timbre variations during acceleration. Several methods can be used to elucidate such timbre variations. In this study, a sensory analysis similar to the one used for car-door noise was first performed on various motor noises (Roussarie et al. 2004). Among the descriptors that were identified by the panel, three were considered essential to characterize the acceleration: "ON" (pronounced [ɔ̃]) characterizes the booming of the motor determined by the audibility of low order, even harmonics; "REU" (pronounced [rœ]) characterizes the roughness of the sound; and “AN” (pronounced [ã]) translates the spectral richness of the motor noise provided by an increased intensity of the odd harmonics. It was hypothesized that the transition between ON and AN was linked to an increased impression of sportiness compared to a monotonous ON sound.

In addition to standard Fourier analyses (Caetano, Saitis, and Siedenburg, Chap. 11), an auditory model that focuses on the perceptually relevant parts of the motor noise was applied to the sound stimuli (Pressnitzer and Gnansia 2005). This model revealed an energy transfer from one group of harmonics toward another during acceleration (Fig. 13.4). To investigate the dynamic aspect of this energy transfer more thoroughly, vocal imitations in which subjects were asked to imitate an accel-

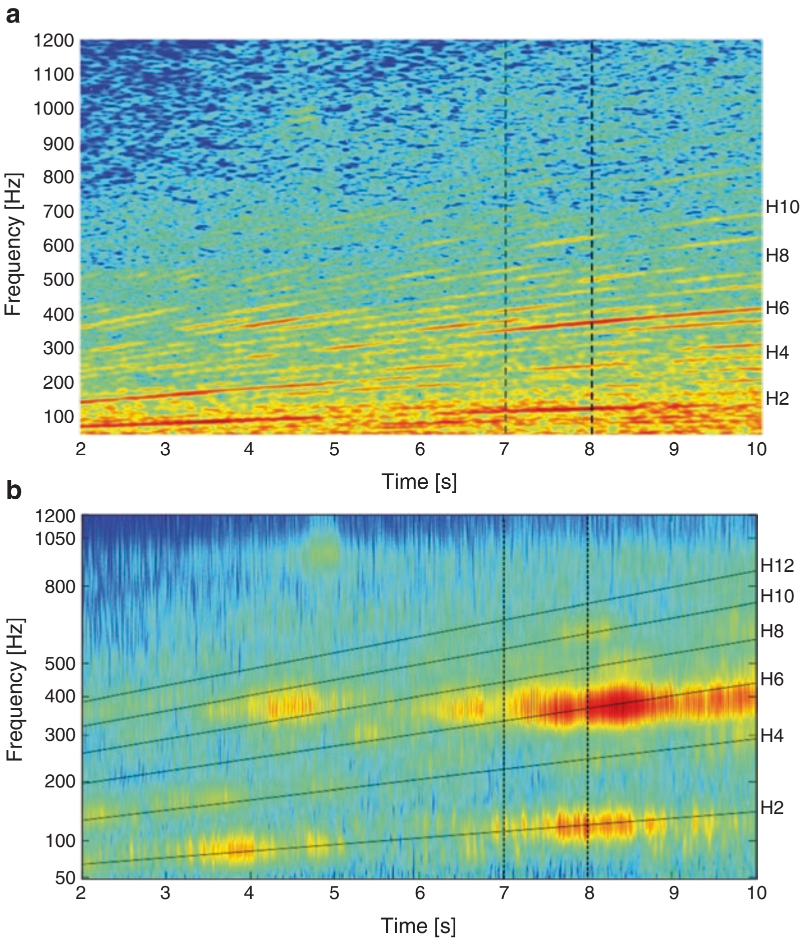

Fig. 13.4 Spectrogram of engine noise during an increase in engine speed (**A**) and cochleogram of the same engine noise (**B**). In the lower part of the figure harmonics are indicated by *solid lines*. The *dotted lines* at 7 s and 8 s indicate a *beating effect* between 300 Hz and 400 Hz revealed by the cochleogram. (**A** from figure 7.4 and **B** from figure 7.5 in Sciabica 2011; used with permission)

erating car were performed. Such approaches have been used, for instance, in the synthesis of guitar sounds using vocal imitation (Traube and Depalle 2004) and to extract relevant features of kitchen sounds (Lemaitre et al. 2011; Lemaitre and Susini, Chap. 9). Vocal imitations are currently used to identify perceptually relevant sound structures of evoked movements and materials (Bordonné et al. 2017).

The vocal imitations provided a simplified description of the dynamic evolution that established a link between the perceived sportiness and the ON/AN transitions. A *source/filter synthesis method* (which consists of filtering an input signal that generally is either a noise or a pulse train) was then developed to control both the roughness (determined by the source) and the formant structure of the sound (determined by the filter). The perceived sportiness could then be controlled intuitively by varying the characteristics of the filter to simulate the ON/AN transition and by modifying the density of the pulse train that constitutes the source related to the roughness (Sciabica et al. 2010, 2012).

13.4.2 Proposing a New Action/Object Paradigm for Sound Synthesis

The last part of this chapter presents the development of new synthesis tools based on perceptual and cognitive studies that unveil perceptually relevant sound morphologies and the construction of synthesis algorithms that are based on these morphologies and thereby enable intuitive sound control. Previous studies on the perception of various sound categories have led to a new sound synthesis paradigm called the *action-object paradigm*, which considers any sound as the result of an action on an object and is based on a semantic description of the sound. This paradigm is coherent with the ecological approach to perception, initially proposed by Gibson (1979) in the case of vision, which suggests the existence of invariant morphological structures associated with the recognition of objects (*structural invariants*) and actions (*transformational invariants*). In this paradigm, the notions of action and object can be considered in a broad sense. Hence, the action can be associated with the dynamics of a sound (temporal evolution) and the object with a sound texture. This paradigm is in line with the phenomenological approach to sound listening adopted by Schaeffer (1966), who proposed a classification system of sounds that he called "typology of sound objects". In this typology, Schaeffer proposes a general classification of sounds (both musical and environmental) that relates to the facture of sounds, that is, the way the energy spreads over time, and the mass related to the spectral content. Facture distinguishes sustained, iterative, and impulsive sounds and can be linked to the perceived action, whereas mass (or spectral content) distinguishes sounds with constant, varying, or indefinable pitch and can be linked to the object (also see Saitis and Weinzierl, Chap. 5).

The action/object paradigm allows for the development of synthesizers that offer sound control from semantic descriptions of the events that created the sound, such as scraping a wooden plate, rubbing a metallic string, or rolling on stones (Conan et al. 2014b). Such synthesizers make it possible to continuously navigate in a sound space based on perceptual invariants of the acoustic signal. This new sound synthesis approach constitutes a radical methodological change and offers new research perspectives in the domains of human perception and cognition, sound design, and musical creation.

13.4.2.1 Perception of Material Categories

A recent study by Aramaki et al. (2011) investigated the perceptual identification of different materials based on impact sounds (see also McAdams, Chap. 2; Agus, Suied, and Pressnitzer, Chap. 3). Particular attention was paid to three different materials: wood, metal, and glass. For this purpose, natural sounds were recorded, analyzed, resynthesized, and tuned to the same pitch class, ignoring octave, to obtain sets of synthetic sounds representative of each material category. A sound morphing process was then applied to obtain sound continua simulating progressive transitions between materials. This morphing process consisted of mixing the spectra between sounds from different material categories and interpolating the damping laws of the two extreme sounds. Each progressive transition between materials was composed of twenty-two hybrid sounds. Participants were asked to categorize all the randomly presented sounds as wood, metal, or glass in a categorization task. Based on the response rates, "typical sounds" were defined as sounds that were classified by more than 70% of the participants in the same material category and "ambiguous sounds" as those that were classified by less than 70% of the participants in a given category. Note that these ambiguous sounds were used in the study presented in Sect. 13.3.2.

While performing the categorization task, reaction times and electrophysiological data were collected using a standard ERP protocol. Analysis of the participants' ERPs showed that the processing of metal sounds differed significantly from the processing of glass and wood sounds as early as 150 ms after the sound onset. These early differences most likely reflect the processing of spectral complexity (Shahin et al. 2005; Kuriki et al. 2006), whereas the later differences observed between the three material categories are likely to reflect differences in sound duration (i.e., differences in damping) (see Alain et al. 2002; McAdams 1999).

The association between the results of the acoustic and electrophysiological analyses suggested that spectral complexity (more precisely the roughness) and both global damping and frequency-dependent damping are relevant cues explaining the perceptual distinction among categories (Aramaki et al. 2011). In particular, both global and frequency-dependent damping differed between categories with metal sounds that had the weakest damping, on the one hand, and sounds from the wood category that were most strongly damped, on the other. Metal sounds also had the largest number of spectral components that introduced roughness in these sounds. Glass sounds had the smallest number of spectral components but a weaker damping than wood sounds.

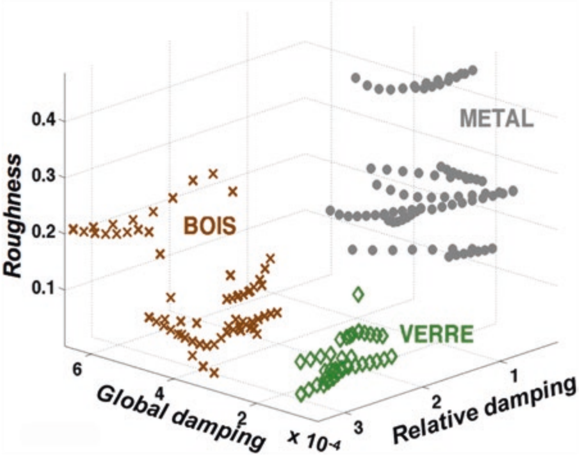

Fig. 13.5 Timbre space of material categories. The distinction between glass, wood, and metal depends on three audio descriptors: global and relative damping and roughness. By acting on these three descriptors continuous transitions between different material categories can be synthesized

These results can be linked to the physical behavior of the sound sources in line with previous studies (see Klatzky et al. 2000; McAdams et al. 2010). The wave propagation process is altered by the characteristics of the medium when the material changes. This process leads to dispersion (due to the stiffness of the material) and dissipation (due to loss mechanisms). Dispersion, which introduces inharmonicity in the spectrum, results from the fact that the wave propagation speed varies depending on the frequency. The dissipation is directly linked to the damping of the sound, which is generally frequency-dependent (high-frequency components are damped more quickly than low-frequency components). These results made it possible to determine the acoustic invariants associated with various sound categories and to propose a timbre space of material categories (Fig. 13.5).

The differences between typical and ambiguous sounds were smaller in the wood-metal and glass-metal continua than in the wood-glass continuum. This is interesting from an acoustic perspective because metal sounds typically present higher spectral complexity (related to the density and repartition of spectral components) than both wood and glass sounds, which have more similar sound properties. Thus, ambiguous sounds in wood-metal and glass-metal continua were easier to categorize than those in the wood-glass continuum, and the ambiguity effect was smaller.

In addition, results showed that ambiguous sounds were associated with slower reaction times than typical sounds. As might be expected, ambiguous sounds are more difficult to categorize than typical sounds. This result is in line with previous findings in the literature showing slower response times for nonmeaningful than for meaningful sounds (e.g., Cummings et al. 2006).

The same categorization protocol was used in a more recent study with participants diagnosed with schizophrenia. The results interestingly revealed that the transitions between material categories were shallower for these participants than for control participants, suggesting the existence of perceptual impairments in such patients due to sensory processing dysfunctions (Micoulaud-Franchi et al. 2011).

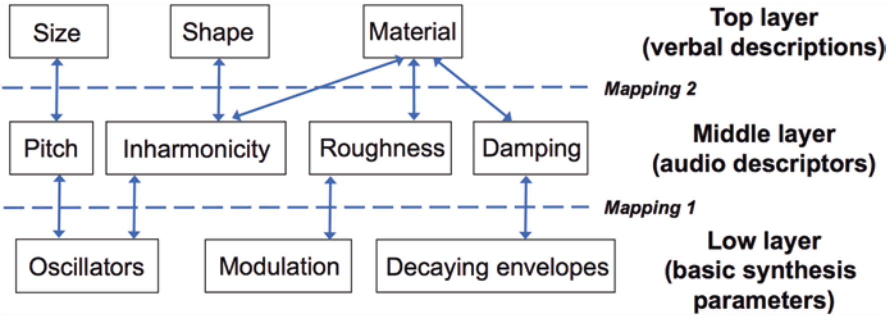

Fig. 13.6 Three-level mapping strategy between basic synthesis parameters (*low layer*), audio descriptors (*middle layer*), and verbal descriptions *top layer*)

The timbre space of material categories is particularly interesting from synthesis and control perspectives since timbre space provides cues for establishing links between low-level synthesis parameters (e.g., amplitudes, frequencies), acoustic descriptors describing pitch and timbre, and semantic labels (wood, metal, glass) that can be used in high-level control interfaces. The positions of the sounds in this material space tell us, for instance, how a sound that evokes wood can be transformed into a sound that evokes metal—by decreasing the damping factors and increasing the roughness. Mapping strategies that enable intuitive controls can therefore be proposed (Fig. 13.6).

13.4.2.2 Perception of Shapes and Excitation

Previous acoustic studies on the links between perception and the physical characteristics of sound sources have brought to light several important properties that can be used to identify the perceived effects of action on an object and the properties of the vibrating object itself (for reviews see Aramaki et al. 2009, 2010). In addition to the frequency-dependent damping and roughness that were found to be important for the perceptual identification of material properties (see McAdams, Chap. 2; Agus, Suied and Pressnitzer, Chap. 3), the perceived hardness of a mallet striking a metallic object is predictable from the characteristics of the attack time.

From a physical point of view, the shape of the impacted object determines the spectral content of the impact sound. The frequencies of the spectral components correspond to the so-called eigenfrequencies, which characterize the modes of the vibrating object and convey important perceptual information about the shape. Previous studies have investigated the auditory perception of physical attributes linked to shape, hollowness, or material. In particular, studies on the geometry of objects have demonstrated that height-width ratios and lengths could be recovered from sounds with reliable accuracy (Lakatos et al. 1997; Carello et al. 1998). Lutfi (2001) showed that the perception of hollowness could be related to frequency judgements and to some extent (depending on the subjects) to acoustic parameters such as damping. Rocchesso (2001) revealed that spherical cavities sounded brighter than cubic cavities, since sounds were more strongly absorbed in cubes than in spheres.

Rakovec et al. (2013) investigated the perception of shapes and found that sounds obtained from striking three-dimensional shapes (e.g., bowls, tubes) were easier to recognize than one-dimensional objects (e.g., bars, strings). The hollow and solid attributes appeared to be quite evocative since no confusion between hollow and solid occurred. The results also revealed a mutual influence between the perceived material and the perceived shape, in line with Tucker and Brown (2002) and Giordano (2003), who found that shape recognition abilities were limited and strongly depended on the material composition.

The perception of the size of the object is mainly correlated with pitch: large objects generally vibrate at lower eigenfrequencies than do small ones. In the case of quasi-harmonic sounds, we assume the pitch to be related to the frequency of the first spectral component. All of these observations lead to the hypothesis of a three-layer mapping strategy that links basic signal parameters via acoustic descriptors to high-level semantic control parameters (see Fig. 13.6).

13.4.2.3 Perception of Various Actions

Invariant sound structures can be linked to the evoked object, such as combinations of specific damping factors and spectral density in the case of material perception, pitch in the case of size perception, or harmonic structure in the case of shape perception. The next question concerns whether invariants linked to the sound-producing action, called transformational invariants, can be found. For this purpose, several studies on continuous interactions between solid objects were performed by considering a subset of continuous interaction sounds: rubbing, scratching, and rolling. Synthesis models for such sounds have already been proposed in previous studies. Some are based on physical modeling (Houben 2002; Stoelinga and Chaigne 2007) or physically informed considerations (van den Doel et al. 2001; Rath and Rocchesso 2004). Others are based on analysis-synthesis schemes (Lagrange et al. 2010; Lee et al. 2010).

Like the control space described in Sect. 13.4.2.1 that allows the user to control the perceived material and to morph continuously from one material to another (e.g., from glass to metal through a continuum of ambiguous materials), a control space that enables continuous control of evoked interactions was developed in this study (e.g., being able to synthesize a rubbing sound and slowly transform it into a rolling one). For this purpose, Conan et al. (2014a) first identified invariants related to the auditory perception of interactions. Phenomenological considerations, physical modeling, and qualitative signal analysis were investigated. They concluded that the interaction forces conveyed the relevant perceptual information regarding the type of interaction. In particular, the interaction force associated with rubbing and scratching sounds could be modeled as an impact series in which impacts are separated by shorter time intervals for rubbing than for scratching. To evoke rolling sounds, it is necessary to consider the strong correlation between the amplitudes of each impact and the time interval that separates them. These findings led to the design of a generic synthesis model that sought to reproduce those interaction forces. An intuitive control space was designed that enables continuous transitions between those interactions.

13.4.3 The Metaphorical Sound Synthesizer

By combining this “action space” with the simulation of the properties of the object (material and shape as described in Sects. 13.4.2.1 and 13.4.2.2), an interface that enables intuitive and interactive real-time control of evoked actions and objects could be designed, as illustrated in Fig. 13.7. This interface offers a large field of sound investigations in which the verbal descriptions of actions and objects only constitute intuitive support to the expression of the composer’s imagination. It should be mentioned that other invariant structures that are not described in this chapter have been identified in the case of evoked motion (Merer et al. 2013), and a synthesizer of environmental sounds offering intuitive control of auditory scenes (rain, waves, wind, fire, footsteps) has been developed by Verron and collaborators (2010). Even if the action/object approach is naturally adapted to the control of realistic sounds produced by objects belonging to our surroundings, one might wonder if such a tool also could satisfy Varèse’s old dream about the creation of “a whole new world of unsuspected sounds”. Hence, the unexpected association between objects and actions might be a means to guide composers in their search for unsuspected or unheard sounds.

The sound space dedicated to the intuitive control of solid objects and their interactions presented earlier in this chapter makes it possible to freely associate actions and objects. This means that it is possible to simulate physically impossible

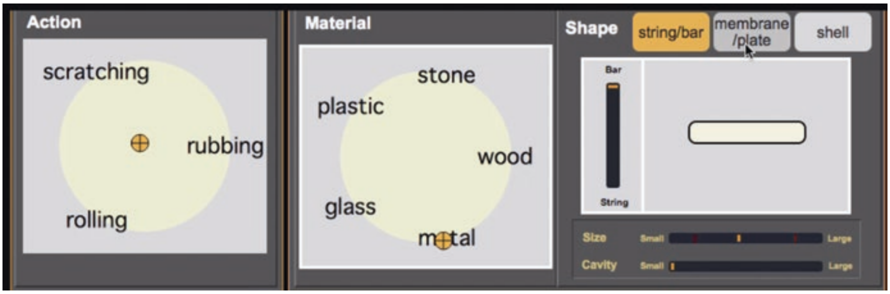

Fig. 13.7 Synthesis interface that enables intuitive and interactive real-time control of evoked actions and objects. The *action* control in the left part enables generation of sounds that evoke different sound-producing actions (e.g., rubbing, scratching); the *material* control in the middle panel evokes different material categories (e.g., wood, metal); the *shape* control in the right panel evokes different shapes (e.g., membrane, string, size)

situations and, for instance, to rub the wind, make a water drop bounce, or make an orchestra squeak. Even if it is difficult to describe sounds that we imagine with words from our language or with metaphoric expressions, new experiences reveal that such a control space opens the door to the creation of unsuspected sounds that conserve the cognitive references of objects and actions due to the invariant structures on which the control space is founded.

In addition to the navigation in this action space, the gesture can be taken into account in the control strategy. Indeed, for such continuous interactions, the underlying gesture is a fundamental attribute that can be conveyed in the dynamics of the sound (Merer et al. 2013; Thoret et al. 2014). Following the synthesis process discussed by van den Doel et al. (2001), the resulting interaction force is low-pass filtered with a cutoff frequency directly related to the relative transversal velocity between the objects that interact (e.g., hand, plectrum) and the surface. When associated with a biological law, a specific calibration of the velocity profile enables the evocation of a human gesture (Thoret et al. 2016). Such a synthesis tool has been used in several applications, for instance, for video games (Pruvost et al. 2015), and can be associated with a graphic tablet to sonify handwriting as a remediation device for dysgraphic children (Danna et al. 2015).

13.5 Summary

The analysis-by-synthesis approach can be used to identify timbre and, more generally, perceptually relevant sound morphologies. Frequency-dependent damping and spectral bandwidth were the most salient descriptors used by xylophone makers when choosing optimal wood species, indicating that they selected highly resonant and crystal-clear sounds. Investigations into the role of timbre in musical performance revealed that musicians consciously used timbre variations to enhance expressiveness. In addition, a detailed study of cello performances showed that ancillary gestures were important and produced a round (as opposed to harsh) timbre.

In the second part of the chapter, particular attention was given to environmental sounds, both in order to better understand how meaning is conveyed by such sounds and to extract sound morphologies that enable the construction of synthesizers that offer easy, intuitive, and continuous sound controls. Electrophysiological measurements have been conducted to investigate how sense is attributed to environmental sounds and to determine whether the brain activity associated with the interpretation of such sounds is similar to the brain activity observed in the case of language processing. Those studies confirmed the existence of a semiotics of isolated sounds, thus suggesting that a language of sounds might be drawn up based on invariant sound structures.

In the last part of the chapter, perceptually salient signal morphologies were identified and associated with evocations of quality, solidity, and sportiness for sounds produced by cars. Invariant structures linked to evocations of solid sounds and their interactions could also be extracted and used to develop synthesizers that enable intuitive control from semantic descriptions of sound events. These developments open the way to new and captivating possibilities for using nonlinguistic sounds for communication. New approaches linked to machine learning (see Caetano, Saitis, and Siedenburg, Chap. 11) and neural responses obtained for spectrotemporal receptive fields (see Elhilali, Chap. 12) should facilitate the development of new tools for generating sound metaphors (see Saitis and Weinzierl, Chap. 5) based on invariant signal structures that can be used to evoke specific mental images via selected perceptual and cognitive attributes. These metaphors can be constructed from scratch or obtained by shaping initially inert sound textures using intuitive (high-level) control approaches.

Compliance with Ethics Requirements Sølvi Ystad declares that she has no conflict of interest.
Mitsuko Aramaki declares that she has no conflict of interest.
Richard Kronland-Martinet declares that he has no conflict of interest.

References

Alain C, Schuler BM, McDonald KL (2002) Neural activity associated with distinguishing concurrent auditory objects. J Acoust Soc Am 111(2):990–995

Angster J, Rucz P, Miklós A (2017) Acoustics of organ pipes and future trends in the research. Acoust Today 13(1):10–18. Spring

Aramaki M, Baillères H, Brancheriau L et al (2007) Sound quality assessment of wood for xylophone bars. J Acoust Soc Am 121(4):2407–2420

Aramaki M, Marie C, Kronland-Martinet R et al (2009) Sound categorization and conceptual priming for non linguistic and linguistic sounds. J Cogn Neurosci 22(11):2555–2569

Aramaki M, Gondre C, Kronland-Martinet R et al (2010) Imagine the sounds: an intuitive control of an impact sound synthesizer. In: Ystad S et al (eds) Auditory Display, LNCS 5954 Springer Heidelberg, p 408–422

Aramaki M, Besson M, Kronland-Martinet R et al (2011) Controlling the perceived material in an impact sound synthesizer. IEEE Trans Audio Speech Lang Process 19(2):301–314

Barthet M, Guillemain P, Kronland-Martinet R et al (2010a) From clarinet control to timbre perception. Acta Acust United Ac 96:678–689

Barthet M, Depalle P, Kronland-Martinet R et al (2010b) Acoustical correlates of timbre and expressiveness in clarinet performance. Music Percept 28(2):135–154

Barthet M, Depalle P, Kronland-Martinet R et al (2011) Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance. Music Percept 28(3):265–279

Bensa J, Bilbao S, Kronland-Martinet R et al (2003) The simulation of piano string vibration: from physical models to finite difference schemes and digital waveguides. J Acoust Soc Am 114(2):1095–1107

Bergman P, Skold A, Vastfjall D et al (2009) Perceptual and emotional categorization of sound. J Acoust Soc Am 126(6):3156–3167

Bezat M C (2007) Perception des bruits d'impact: application au bruit de fermeture de porte automobile ("Perception of impact noises: application to car-door closure noise"). Ph.D. dissertation, Université Aix-Marseille I

Bezat MC, Kronland-Martinet R, Roussarie V et al (2014) From acoustic descriptors to evoked quality of car-door sounds. J Acoust Soc Am 136(1):226–241

Bilbao S, Webb C (2013) Physical modeling of timpani drums in 3D on GPGPUs. J Audio Eng Soc 61(10):737–748

Bordonné T, Dias-Alves M, Aramaki M et al (2017) Assessing sound perception through vocal imitations of sounds that evoke movements and materials. In: Proceedings of the 13th international symposium on computer music multidisciplinary research "Music Technology with Swing", Matosinhos, 25–28 September 2017

Carello C, Anderson KL, Kunkler-Peck AJ (1998) Perception of object length by sound. Psychol Sci 26(1):211–214

Chadefaux D, Le Carrou J-L, Wanderley MM et al (2013) Gestural strategies in the harp performance. Acta Acust United Ac 99(6):986–996

Chaunier L, Courcoux P, Della Valle G et al (2005) Physical and sensory evaluation of cornflakes crispness. J Texture Stud 36:93–118

Conan S, Derrien O, Aramaki M et al (2014a) A synthesis model with intuitive control capabilities for rolling sounds. IEEE Trans Audio Speech Lang Process 22(8):1260–1273

Conan S, Thoret E, Aramaki M et al (2014b) An intuitive synthesizer of continuous interaction sounds: rubbing, scratching and rolling. Comput Music J 38(4):24–37. https://doi.org/10.1162/COMJ_a_00266

Cook P (2001) Principles for designing computer music controllers. In: Proceeding of new interfaces for musical expression, NIME-01, Seattle, 1–2 Apr 2001

Cook P, Scavone G (1999) The synthesis toolkit. In: Proceedings of the international computer music conference, Beijing, 22–27 October 1999

Cummings A, Ceponiene R, Koyama A et al (2006) Auditory semantic networks for words and natural sounds. Brain Res 1115:92–107

Danna J, Paz-Villagrán V, Gondre C et al (2015) Let me hear your handwriting! Evaluating the movement fluency from its sonification. PLoS One 10(6):e0128388

Bedos de Celles D F (1766) L'art du facteur d'orgues. ("The art of the organ builder"). Reprint of the original edition from Paris by Slatkine, 2004, ISBN 2051019398

De Wilde L (2016) Les fous du son, d'Edison à nos jours. Ed. Grasset

Desmet F, Nijs L, Demey M et al (2012) Assessing a clarinet player's performer gestures in relation to locally intended musical targets. J New Music Res 41(1):31–48

Fourier J (1878) The analytical theory of heat. Cambridge University Press, Cambridge

Gaver WW (1993) What in the world do we hear? An ecological approach to auditory event perception. Ecol Psychol 5(1):1–29

Gibson JJ (1979) The ecological approach to visual perception. Houghton Mifflin, Boston

Giordano BL (2003) Material categorization and hardness scaling in real and synthetic impact sounds. In: Rocchesso D, Fontana F (eds) The sounding object. Mondo Estremo, Firenze

Giordano BL, McAdams S (2006) Material identification of real impact sounds: effects of size variation in steel, wood, and plexiglass plates. J Acoust Soc Am 119(2):1171–1181

Gobin P, Kronland-Martinet R, Lagesse GA et al (2003) From sounds to music: different approaches to event piloted instruments. In: Wiil UK (ed) LNCS 2771. Springer, Heidelberg, pp 225–246

Goodchild M, Wild J, McAdams S (2017) Exploring emotional responses to orchestral gestures. Music Sci:1–25. https://doi.org/10.1177/1029864917704033

Grey JM (1977) Multidimensional perceptual scaling of musical timbres. J Acoust Soc Am 61:1270–1277

Guillemain P, Kergomard J, Voinier T (2005) Real-time synthesis of clarinet-like instruments using digital impedance models. J Acoust Soc Am 118:483–494

Helmholtz H (1868) On the sensations of tone as a physiological basis for the theory of music. Longmans, Green, and co, New York

Houben M (2002) The sound of rolling objects, perception of size and speed. PhD dissertation, Technische Universiteit, Eindhoven

Huang NE, Shen Z, Long S et al (1998) The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis. Proc R Soc A Lond 454(1971):903–995

Jensenius AR (2007) Action-Sound. Developing Methods and Tools to Study Music-Related Body Movement. PhD dissertation, University of Oslo

Klatzky RL, Pai DK, Krotkov EP (2000) Perception of material from contact sounds. Presence Teleop Virt 9(4):399–410

Krimphoff J, McAdams S, Winsberg S (1994) Caractérisation du timbre des sons complexes, II Analyses acoustiques et quantification psychophysique (Characterization of complex sounds timbre, II Acoustical analyses and psychophysical quantification). J Physique IV, Colloque C5, 4:625–628

Kronland-Martinet R, Guillemain P, Ystad S (1997) Modelling of natural sounds using time-frequency and wavelet representations. Organised sound,. Cambridge University Press 2(3):179–191

Kuriki S, Kanda S, Hirata Y (2006) Effects of musical experience on different components of meg responses elicited by sequential piano-tones and chords. J Neurosci 26(15):4046–4053

Kutas M, Hillyard SA (1980) Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207:203–204

Kuwano S, Fastl H, Namba S et al (2006) Quality of door sounds of passenger cars. Acoust Sci Technol 27(5):309–312

Lagrange M, Scavone G, Depalle P (2010) Analysis/synthesis of sounds generated by sustained contact between rigid objects. IEEE Trans Audio Speech Lang Process 18(3):509–518

Lakatos S, MacAdams S, Caussé R (1997) The representation of auditory source characteristics: simple geometric form. Atten Percept Psychophys 59(8):1180–1190

Lee J S, Depalle P, Scavone G (2010) Analysis/synthesis of rolling sounds using a source-filter approach. In: 13th international conference on digital audio effects (DAFx-10), Graz, 6–10 Sept 2010

Lemaitre G, Susini P, Winsberg S et al (2009) The sound quality of car horns : designing new representative sound. Acta Acustica United Ac 95(2)

Lemaitre G, Dessein A, Susini P et al (2011) Vocal imitations and the identification of sound events. Ecol Psychol 4(23):267–307

Lembke SA, Levine S, McAdams S (2017a) Blending between bassoon and horn players: an analysis of timbral adjustments during musical performance. Music Percept 35(2):144–164

Lembke S-A, Parker K, Narmour E et al (2017b) Acoustical correlates of perceptual blend in timbre dyads and triads. Music Sci. https://doi.org/10.1177/1029864917731806

Lutfi RA (2001) Auditory detection of hollowness. J Acoust Soc Am 110(2):1010–1019

Lutfi RA, Oh EL (1997) Auditory discrimination of material changes in a struck-clamped bar. J Acoust Soc Am 102(6):3647–3656

Mathews MV (1963) The digital computer as a musical instrument. Science 142(3592):553–557

Mathews MV, Miller JE, Pierce JR et al (1965) Computer study of violin tones. J Acoust Soc Am 38(5):912–913

McAdams S (1999) Perspectives on the contribution of timbre to musical structure. Comput Music J 23(3):85–102

McAdams S, Bigand E (1993) Thinking in sound: the cognitive psychology of human audition. Oxford University Press, Oxford

McAdams S, Roussarie V, Chaigne A et al (2010) The psychomechanics of simulated sound sources: material properties of impacted thin plates. J Acoust Soc Am 128(3):1401–1413. https://doi.org/10.1121/1.3466867

Merer A, Ystad S, Kronland-Martinet R et al (2011) Abstract sounds and their applications in audio and perception research. In: Ystad et al (eds) Exploring music contents, LNCS vol. 6684, Springer Heidelberg, pp 176–187

Merer A, Aramaki M, Ystad S et al (2013) Perceptual characterization of motion evoked by sounds for synthesis control purposes. ACM Trans Appl Percept 10(1):1–24

Micoulaud-Franchi JA, Aramaki M, Merer A et al (2011) Categorization and timbre perception of environmental sounds in schizophrenia. Psychiatry Res 189(1):149–152

Miranda R, Wanderley MM (2006) New digital musical instruments: control and interaction beyond the keyboard. A-R Editions, Middleton

Moog R (1987) Position and force sensors and their application to keyboards and related controllers. In: Proceedings of the AES 5th international conference: music and digital technology, New York, May 1987

Moore TR (2016) The acoustics of brass musical instruments. Acoust Today 12(4):30–37. Winter

Peeters G, Giordano B, Susini P et al (2011) The timbre toolbox: audio descriptors of musical signals. J Acoust Soc Am 130(5):2902–2916

Pierce J (1965) Portrait of the machine as a young artist. Playboy 12(6):124–125. 150, 182, 184

Pressnitzer D, Gnansia D (2005) Real-time auditory models. In: Proceedings of the international computer music conference, Barcelona, Spain, 5–9 Sept, 2005

Pruvost L, Scherrer B, Aramaki M, et al (2015) Perception-based interactive sound synthesis of morphing solids' interactions. In: Proceedings of the Siggraph Asia 2015, Kobe, 2–5 Nov 2015

Rakovec C E, Aramaki M, Kronland-Martinet R (2013) Perception of material and shape of impacted everyday objects. In: Proceedings of the 10th international symposium on computer music multidisciplinary research, Marseille, 15–18 Oct 2013

Rath M, Rocchesso D (2004) Continuous sonic feedback from a rolling ball. IEEE MultiMedia 12(2):60–69

Risset JC (1965) Computer study of trumpet sounds. J Acoust Soc Am 38(5):912

Roads C (1988) Introduction to Granular Synthesis. Comput Music J 12(2):11–13. https://doi.org/10.2307/3679937

Rocchesso D (2001) Acoustic cues for 3-D shape information. In: Proceedings of the 2001 international conference on auditory display, Espoo, 29 July-1 Aug 2001

Roussarie V (2005) What's so hot About sound? – Influence of HVAC sound on thermal comfort. In: Proceedings of the inter-noise, Rio de Janeiro, 7–10 Aug 2005

Roussarie V, Richard F, Bezat M C (2004) Validation of auditory attributes using analysis synthesis method. In: Proceedings of the CFA/DAGA'04, Strasbourg, 22–25 Mar 2004

Rozé J, Aramaki M, Kronland-Martinet R et al (2016) Exploring the effects of constraints on the cellist's postural displacements and their musical expressivity. In: Kronland-Martinet R et al (eds) Music, mind and embodiment, LNCS 9617. Springer, Heidelberg, pp 22–41

Rozé J, Aramaki M, Kronland-Martinet R et al (2017) Exploring the perceived harshness of cello sounds by morphing and synthesis techniques. J Acoust Soc Am 141(3):2121–2136

Saitis C, Giordano B, Fritz C et al (2012) Perceptual evaluation of violins: a quantitative analysis of preference judgments by experienced players. J Acoust Soc Am 132(6):4002–4012

Saitis C, Fritz C, Scavone G et al (2017) A psycholinguistic analysis of preference verbal descriptors by experienced musicians. J Acoust Soc Am 141(4):2746–2757

Saitis C, Järveläinen H, Fritz C (2018) The role of haptic cues in musical instrument quality perception. In: Papetti S, Saitis C (eds) Musical haptics. Springer, Cham, pp 73–93

Schaeffer P (1966) Traité des objets musicaux. Éditions du Seuil, Paris

Schaeffer P (2017) Treatise on musical objects: an essay across disciplines (trans: North C, Dack J). University of California Press, Oakland

Scholl D, Amman S (1999) A new wavelet technique for transient sound visualization and application to automotive door closing events. In: Proceedings of the SAE noise and vibration conference and exposition, Traverse City, MI

Schön D, Ystad S, Kronland-Martinet R et al (2009) The evocative power of sounds: conceptual priming between words and nonverbal sounds. J Cogn Neurosci 22:1026–1035

Sciabica J F (2011) Modélisation perceptive du bruit habitacle et caractérisation de son ressenti. (Perceptual modelling of interior car noise and characterization of induced sensation), Ph.D. dissertation, Aix-Marseille Université

Sciabica JF, Bezat MC, Roussarie V et al (2010) Timbre characteristics of interior car sound. In: Ystad S et al (eds) Auditory display. Springer, Heidelberg, pp 377–391

Sciabica J F, Olivero A, Roussarie V, et al (2012) Dissimilarity test modelisation applied to interior car sound perception. In: Proceedings of the 45th international conference on applications of time-frequency processing in audio, Helsinki, 1–4 Mar 2012

Shahin A, Roberts LE, Pantev C et al (2005) Modulation of p2 auditory-evoked responses by the spectral complexity of musical sounds. Neuroreport 16(16):1781–1785

Smalley D (1994) Defining Timbre – Refining Timbre. Contemp Music Rev 10:35–48

Stoelinga C, Chaigne A (2007) Time-domain modeling and simulation of rolling objects. Acta Acust United Ac 93(2):290–304

Suied C, Susini P, McAdams S (2008) Evaluating warning sound urgency with reaction times. J Exp Psychol Appl 14:201–212

Sundberg J (2000) Grouping and differentiation two main principles in the performance of music. In: Nakada T (ed) Integrated human brain science: theory, method application (Music). Elsevier Science B.V, Amsterdam

Susini P, McAdams S, Winsberg S et al (2004) Characterizing the sound quality of air-conditioning noise. Appl Acoust 65(8):763–790

Thompson MR, Luck G (2012) Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. Music Sci 16(1):19–40

Thoret E, Aramaki M, Kronland-Martinet R, Velay JL, Ystad S (2014) From sound to shape: Auditory perception of drawing movements. J Exp Psychol Hum Percept Perform 40(3): 983–994. https://doi.org/10.1037/a0035441

Thoret E, Aramaki M, Bringoux L, Ystad S, Kronland-Martinet R (2016) Seeing circles and drawing ellipses: when sound biases reproduction of visual motion. PLoS ONE 11(4):e0154475. https://doi.org/10.1371/journal.pone.0154475

Traube C, Depalle P (2004) Timbral analogies between vowels and plucked strings tones. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP04) Montreal, 17–21 May 2004

Tucker S, Brown G J (2002) Investigating the perception of the size, shape and mateial of damped and free vibrating plates. Technical Report, Department of Computer Science, University of Sheffield

van den Doel K, Kry P G, Pai, D K (2001) FoleyAutomatic: physically-based sound effects for interactive simulation and animation. In: Proceedings of the 28th annual conference on computer graphics and interactive techniques. ACM, Los Angeles, 12–17 Aug 2001

Van Zijl AG, Luck G (2013) Moved through music: the effect of experienced emotions on performers' movement characteristics. Psychol Music 41(2):175–197

Varèse E (1917) Que la Musique Sonne. 391, n°5, June 1917

Verron C, Aramaki M, Kronland-Martinet R et al (2010) A 3D immersive synthesizer for environmental sounds. IEEE Trans Audio Speech Lang Process 18(6):1550–1561

Wanderley MM, Vines BW, Middleton N et al (2005) The musical significance of clarinetists' ancillary gestures: an exploration of the field. J New Music Res 34(1):97–113

Warren W, Verbrugge R (1984) Auditory perception of breaking and bouncing events: a case study in ecological acoustics. J Exp Psychol Hum Percept Perform 10(5):704–712

Wessel DL (1979) Timbre space as a musical control structure. Comput Music J 3(2):45–52

Wildes RP, Richards WA (1988) Recovering material properties from sound. MIT Press, Cambridge, MA, pp 356–363

Wolfe J (2018) The acoustics of woodwind musical instruments. Acoustics Today,. Spring 2018 14(1):50–56