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 bio-cognitive 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