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 con-

cepts, 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.

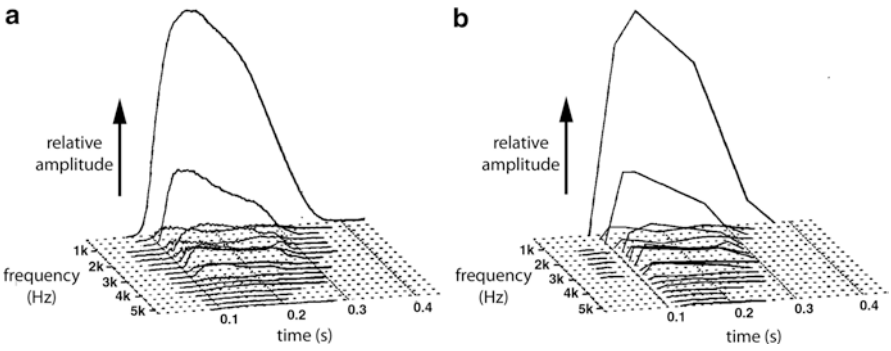

Fig. 2.1 Analysis of the time-varying amplitudes and frequencies of the partials of a bass clarinet tone (**a**) and their simplification by line segment functions (**b**). In this three-dimensional representation, *time* goes from left to right, *relative amplitude* from bottom to top, and *frequency* 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

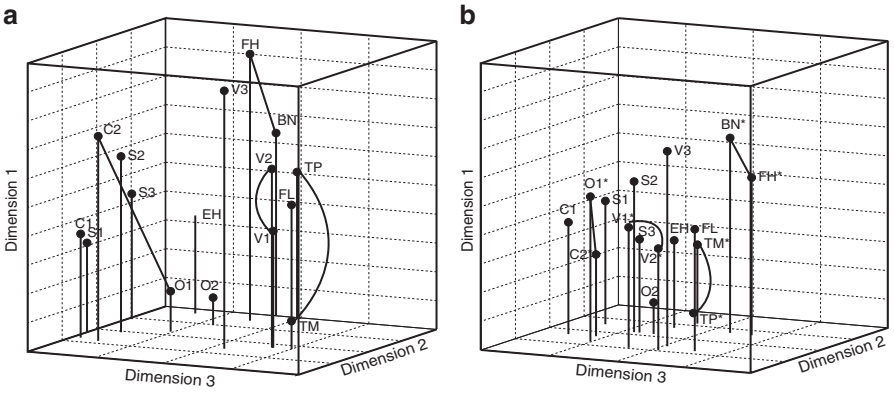

Fig. 2.2 Timbre spaces: (**a**) from Grey (1977) and (**b**) from Grey and Gordon (1978). *Dimension 1* is related to the spectral envelope distribution. *Dimension 2* corresponds to the amount of fluctuation over time in the spectral envelope and the synchrony of onset of the harmonics. *Dimension 3* 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

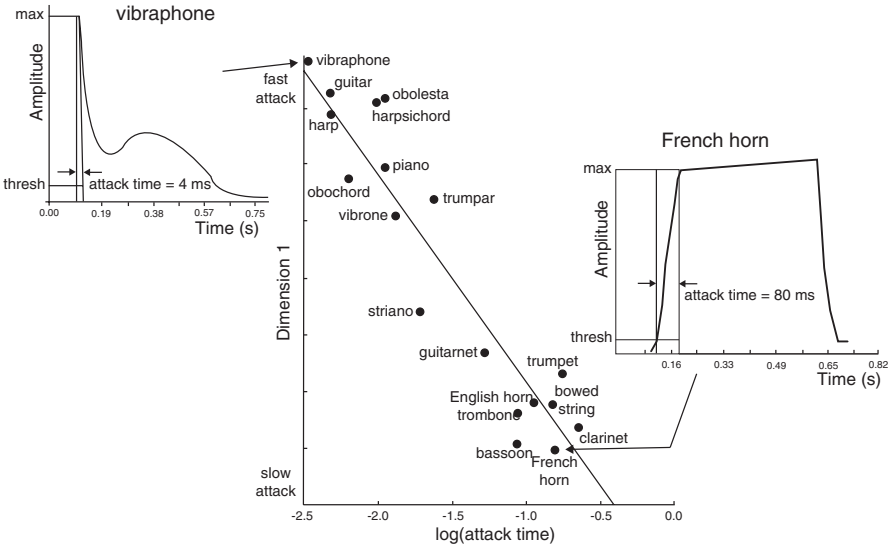

Fig. 2.3 Relationship of log (attack time) to position along Dimension 1. The diagrams for vibraphone and French horn show the *global amplitude envelope* (amplitude variation over time). The *attack time* was measured as the time from a threshold value to the maximum in the amplitude envelope. The attack 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)

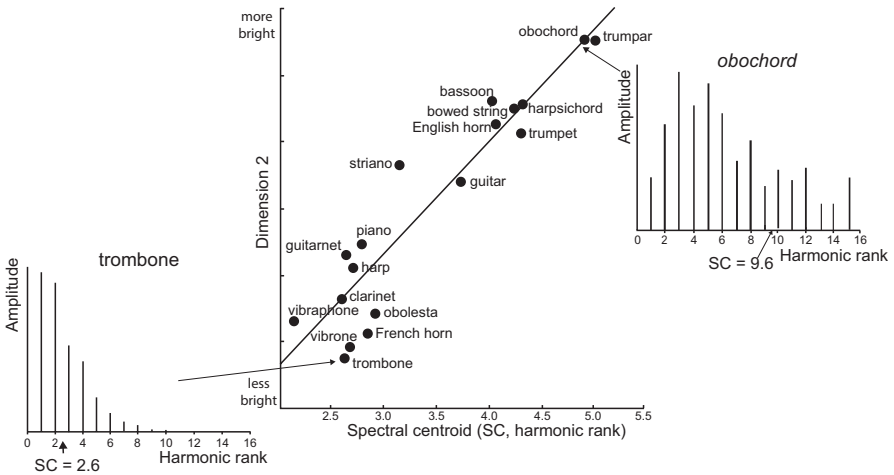

Fig. 2.4 Relationship of the spectral centroid to position along Dimension 2. The diagrams for *trombone* and *obochord* show the frequency spectra. The balance point in the energy spectrum (*spectral centroid, SC*) for each is shown. The SC is lower for trombone, which has much less energy in the higher 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)

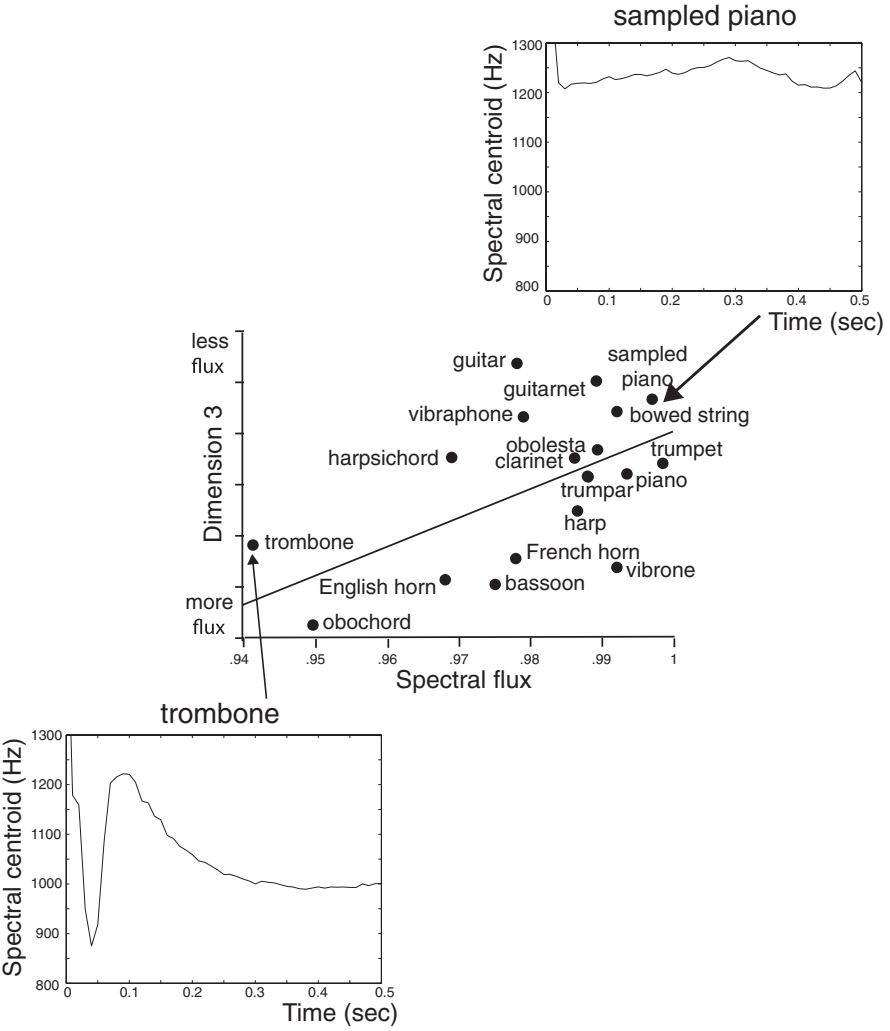

Fig. 2.5 Relationship of spectral flux to position along Dimension 3. The measure of *spectral flux* is the average correlation of the spectra between adjacent time frames. *Less flux* gives higher correlation values and *more flux* gives lower values. The diagrams for trombone and sampled piano show 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

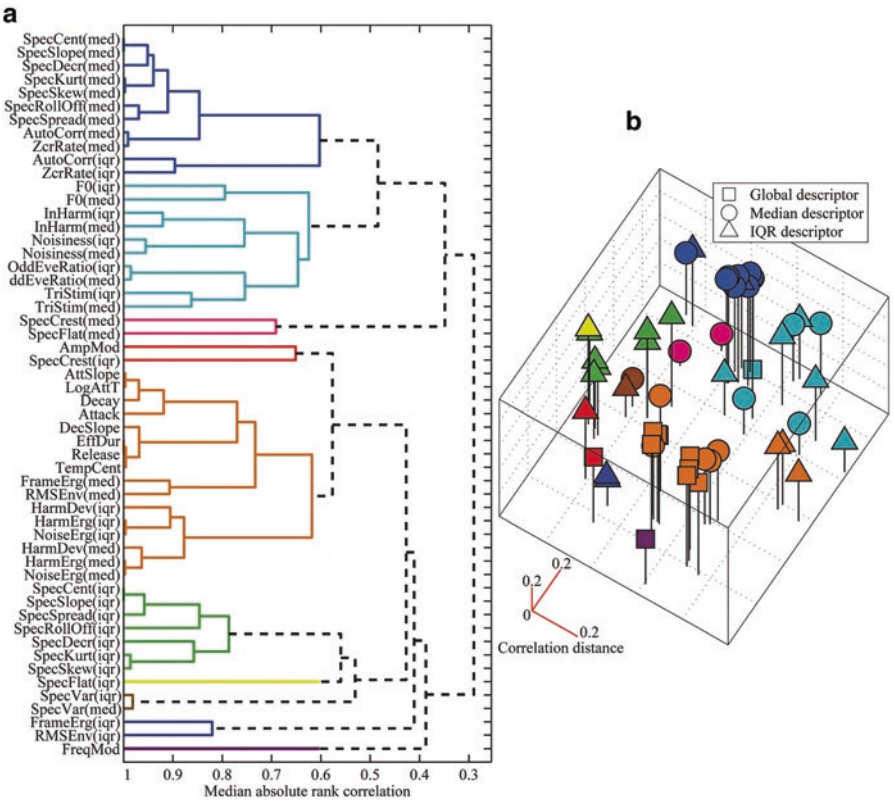

Fig. 2.6 Structure of similarities among audio descriptors. (**a**) The results of a hierarchical cluster analysis of the correlations among the audio descriptors listed along the y axis. Scalar values derived from the temporal energy envelope cluster in the middle. Statistical measures of time-varying 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 inti-

mately 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).

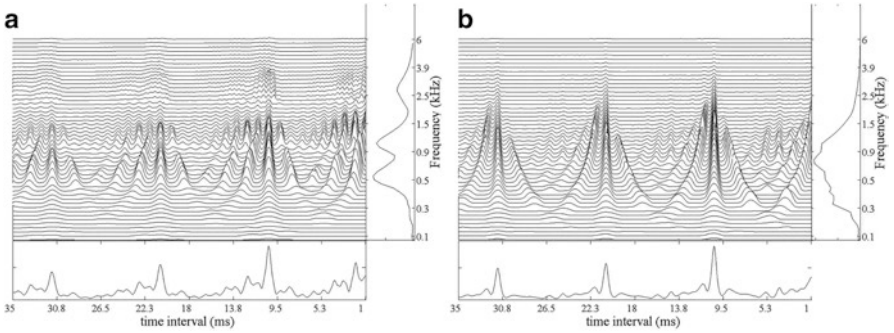

Fig. 2.7 Simulated auditory images of sustained parts of tones produced by a baritone voice **(a)** and a French horn **(b)** at the same fundamental frequency. Each line in the auditory image shows the simulated activity in a given frequency channel (auditory filters) over time. The *lower diagrams* 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.

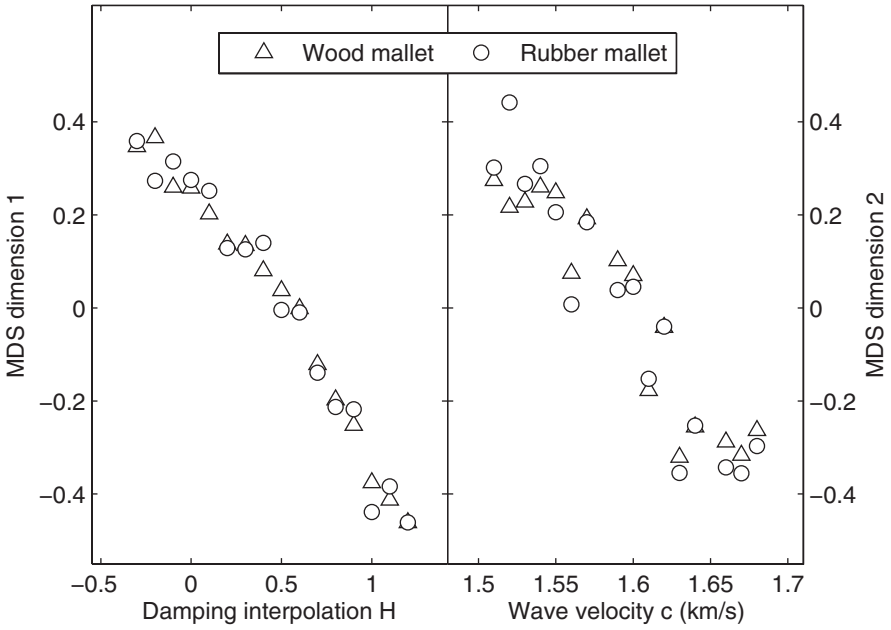

Fig. 2.8 Relation between MDS dimensions and physical model parameters. In the **left panel**, the coordinate along Dimension 1 is plotted as a function of the factor that controlled the interpolation between viscoelastic and thermoelastic damping (*damping interpolation H*). In the **right panel**, 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

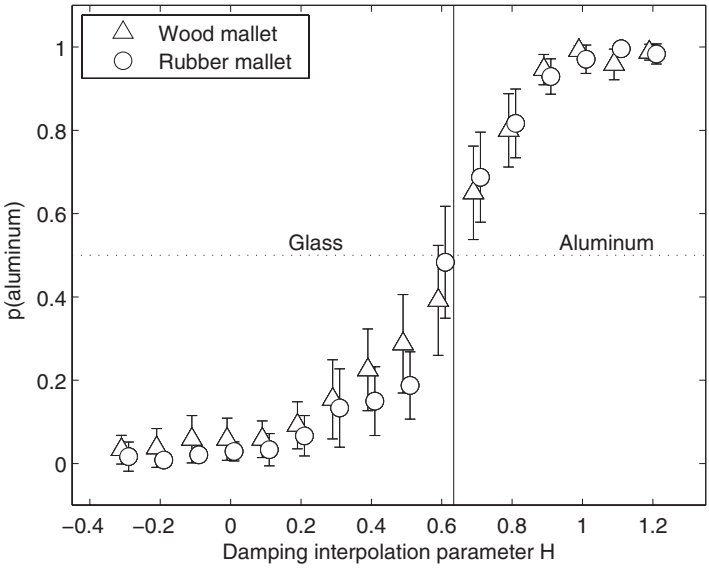

Fig. 2.9 Probability of classifying a sound as aluminum. Error bars are 95% confidence intervals. The probability (*p*) that a sound was classified as aluminum is plotted as a function of the damping interpolation parameter (*H*). The *vertical line* indicates the point of inflection of the curve and 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

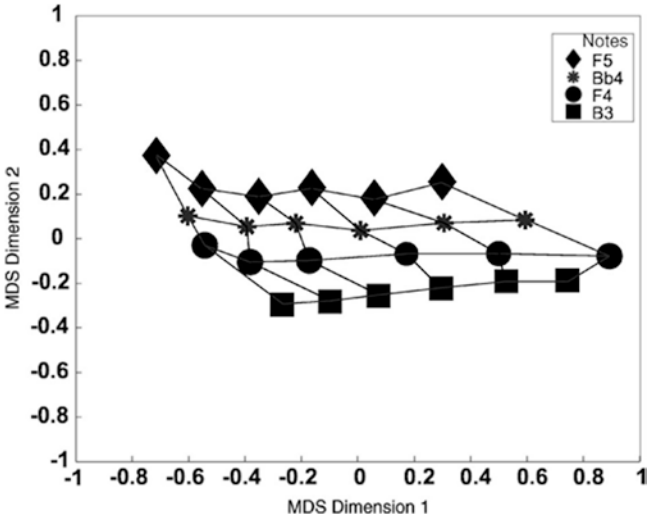

Fig. 2.10 Multidimensional scaling (MDS) solution in two dimensions rotated to maximize correlation between Dimension 1 and the spectral centroid. Note that the musical notes at different fundamental frequencies (different symbols) are not strongly affected by spectral centroid: the curves are flat. 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 dis-

crimination 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