Chapter 11
Audio Content Descriptors of Timbre

Marcelo Caetano, Charalampos Saitis, and Kai Siedenburg

Abstract This chapter introduces acoustic modeling of timbre with the audio descriptors commonly used in music, speech, and environmental sound studies. These descriptors derive from different representations of sound, ranging from the waveform to sophisticated time-frequency transforms. Each representation is more appropriate for a specific aspect of sound description that is dependent on the information captured. Auditory models of both temporal and spectral information can be related to aspects of timbre perception, whereas the excitation-filter model of sound production provides links to the acoustics of sound production. A brief review of the most common representations of audio signals used to extract audio descriptors related to timbre is followed by a discussion of the audio descriptor extraction process using those representations. This chapter covers traditional temporal and spectral descriptors, including harmonic description, time-varying descriptors, and techniques for descriptor selection and descriptor decomposition. The discussion is focused on conceptual aspects of the acoustic modeling of timbre and the relationship between the descriptors and timbre perception, semantics, and cognition, including illustrative examples. The applications covered in this chapter range from timbre psychoacoustics and multimedia descriptions to computer-aided orchestra-

Electronic supplementary material: The online version of this chapter (https://doi.org/10.1007/978-3-030-14832-4_11) contains supplementary material, which is available to authorized users.

M. Caetano (✉)
Sound and Music Computing Group, INESC TEC, Porto, Portugal
e-mail: mcaetano@inesctec.pt

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

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

K. Siedenburg et al. (eds.), *Timbre: Acoustics, Perception, and Cognition*,
Springer Handbook of Auditory Research 69,
https://doi.org/10.1007/978-3-030-14832-4_11

tion and sound morphing. Finally, the chapter concludes with speculation on the role of deep learning in the future of timbre description and on the challenges of audio content descriptors of timbre.

Keywords Environmental sound · Excitation-filter model · Machine learning · Musical instrument · Pattern recognition · Sound color · Speech · Time-frequency analysis

11.1 Introduction

A sound wave carries a pattern of oscillations, which was generated by a driving force that excited a vibrating object through a physical medium such as the air. When the sound wave reaches the ear, these oscillations are processed and interpreted by the brain as sound. On its way from the source to the ear, the sound wave carries precise information about the vibrating object (e.g., a cello), the driving force (e.g., bowed), and possibly other physical objects with which it interacted (e.g., the walls of a concert hall). The human brain has a remarkable ability to convert the detailed information contained in sound waves into the meaningful experience of hearing—from the minute inflections of speech that facilitate human communication to the expressiveness of microvariations in music (Handel 1995). But how do sound waves convey identifiable properties of the sound source, of the sound-generating event, and even of the objects with which the sound wave interacted? What aspects of the audio representation of the sound wave, commonly called the waveform, carry information about the size or material of the source, the type of excitation (e.g., knocking or rubbing) that generated it, or its perceived timbre? What is the acoustic basis of perceived dissimilarities, such as those between different instruments, different registers of the same instrument, and different players playing the same instrument? This chapter examines how differences in timbre manifest themselves in the audio signal and how such information can be extracted computationally from different signal representations in the form of *audio descriptors* to acoustically characterize timbre in music, speech, and environmental sounds.

Today timbre is understood from two perceptual viewpoints: as a sensory quality and as a contributor to source identity (Siedenburg and McAdams 2017). In the former, two sounds can be declared qualitatively dissimilar without bearing source-cause associations. In the latter, timbre is seen as the primary perceptual vehicle for the recognition and tracking over time of the identity of a sound source. Both approaches consider timbre as a very complex set of perceptual attributes that are not accounted for by pitch, loudness, duration, spatial position, and spatial characteristics such as room reverberation (Siedenburg, Saitis, and McAdams, Chap. 1). When timbre is viewed as qualia, its attributes underpin dissimilarity (McAdams, Chap. 2) and semantic ratings (Saitis and Weinzierl, Chap. 5). In the timbre-as-identity scenario, they facilitate sound source recognition (Agus, Suied, and Pressnitzer, Chap. 3). Further adding to its complex nature, timbre functions on different scales of detail (Siedenburg and McAdams 2017) in the sense that one sound-producing object can yield multiple distinct timbres (Barthet et al. 2010), and timbres from sound-producing objects of the same type but different "make" may differ substantially enough to affect quality judgements (Saitis et al. 2012). How informative is a given audio descriptor when examining different scales of timbre? What is the acoustic difference between a note played *pianissimo* and the same note played *fortissimo* or notes played in different registers on the same instrument?

Some of the most successful attempts to establish relationships between audio descriptors and perceptual aspects of timbre have resulted from multidimensional scaling (MDS) of pairwise dissimilarity ratings between musical instrument sounds (Grey and Gordon 1978; McAdams et al. 1995). Descriptors calculated from temporal and spectrotemporal representations of the audio signal are typically correlated with the dimensions of MDS *timbre spaces* to capture the acoustic cues underpinning the mental representation of timbre (McAdams, Chap. 2). Beyond psychoacoustics and music psychology, extracting quantitative descriptors potentially related to timbre from audio signals is an important part of the music information retrieval (MIR) discipline (Casey et al. 2008; Levy and Sandler 2009). The MIR task most relevant to timbre per se is musical instrument classification, which relies on an ensemble of descriptors associated with both the excitation-filter model and time-frequency representations to classify musical instrument sounds. However, the way audio descriptors are approached by MIR diverges from psychology due to differences in epistemic traditions and scientific goals between the two disciplines (Siedenburg et al. 2016a), a point discussed further in Sect. 11.4.1.

In MIR, descriptors are more commonly referred to as *features*. In psychology, features are discrete whereas dimensions are continuous (Peeters et al. 2011). In multimedia, features are perceptual by nature and descriptors are representations of features with specific instantiations (i.e., values) associated with data (Nack and Lindsay 1999). Pitch, for instance, is a feature of periodic sounds; fundamental frequency f_0 is a possible descriptor of pitch and $f_0 = 440$ Hz is the corresponding descriptor value. In MIR, features are extracted from the audio independently of the intrinsic nature of the information they represent (Casey et al. 2008). As such, a chord, the melody, and even the spectral envelope can be a feature. Following Peeters et al. (2011), the term *descriptor* is adopted here to disambiguate the concept of extracting information from the audio signal to describe its content.

Key questions that arise in working with audio descriptors of timbre include the following:

- What audio descriptors are appropriate for different tasks?
- What is the relation between the information captured by the descriptor and its usefulness?
- What is the relation between audio descriptors and perceptual, semantic, and cognitive aspects of timbre?
- What temporal information is important for timbre and how should it be represented with descriptors?
- How do we deal with timbral dimensions that covary with other perceptual dimensions?

Attempting to provide some answers to these questions, this chapter lays out a pathway into audio descriptor design and application. Section 11.2 presents basic audio representations that serve as a starting point for the extraction of audio descriptors presented in Sect. 11.3. Subsequently, Sect. 11.4 explores important applications of audio descriptors in the domain of timbre psychoacoustics, sound meta-description, musical orchestration, and sound morphing. Section 11.5 closes with a discussion of deep learning in automatic audio description and promising avenues for future research.

11.2 Representations of Audio Signals

This section introduces basic mathematical representations of audio from which audio descriptors can be extracted, which is not intended as a thorough explanation with a full mathematical treatment but rather as a general intuitive overview of the main concepts involved. The waveform (see Figs. 11.1–11.3) represents the pattern of pressure oscillations of a sound wave. Positive amplitude corresponds to compression and negative amplitude represents rarefaction. In digital audio, the discrete representation of sound waves is obtained by sampling continuous waveforms. Specifically, a discrete waveform is the result of sampling its analog counterpart at regular time intervals. The waveform contains all of the information carried by the sound wave it represents, but the waveform itself is seldom useful as a representation from which to extract perceptually meaningful information about timbral attributes or to categorize the sound source.

Figures 11.1–11.3 illustrate a typical sequence of steps taken to transform a waveform into a representation suitable for audio content description. The waveform is first windowed into time frames and the spectrum of each frame is obtained with the discrete Fourier transform (DFT). Descriptors are then computed globally or for each time frame. Details of the different steps are discussed below and in Sect. 11.3. To hear the sounds used in Figs. 11.1–11.3, go to the sound files “music.mp3”, “speech.mp3”, and “water.mp3”.

11.2.1 Short-Time Fourier Transform and Spectrogram

The DFT is the standard method to obtain a representation of the frequency decomposition of the waveform called the frequency spectrum (Jaffe 1987a, b). The discrete frequencies of the DFT are linearly spaced, which means that adjacent frequency samples are separated by a constant interval called a *frequency bin*.

The short-time Fourier transform (STFT) analyzes a signal in terms of time *and* frequency by viewing it through successive overlapping windows, as depicted in Figs. 11.1–11.3, and then taking the DFT of each windowed frame (Portnoff 1980). The effect of the window is to concentrate the information in a short temporal frame.

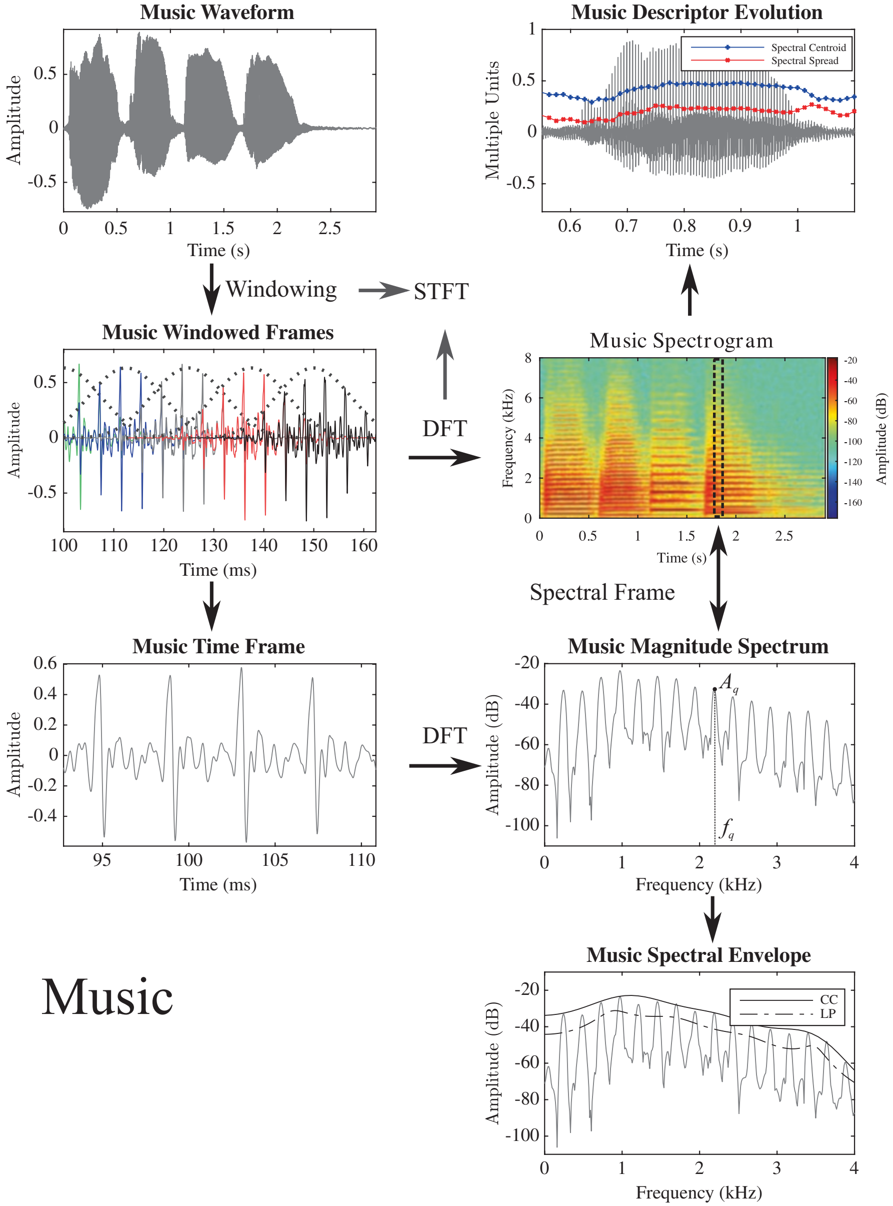

Fig. 11.1 Illustration of the sequence of steps to extract audio content descriptors from music. The *music* excerpt comprises four isolated notes (B3, G3, A4, and D4) of the monophonic trumpet track from a multitrack recording of Beethoven's Symphony No. 5. To hear the sounds, go to the sound file "music.mp3". The arrows indicate the connections between the panels. The extraction of descriptors from time-frequency representations is illustrated going *counter-clockwise* from the panel labeled *waveform* and the extraction of descriptors from the excitation-filter model is illustrated going *clockwise* from the panel labeled *time frame*. Abbreviations: A_q, amplitude of qth partial; *CC*, cepstral coefficients; *DFT*, discrete Fourier transform; f_q, frequency of qth partial; *LP*, linear prediction; *STFT*, short-time Fourier transform

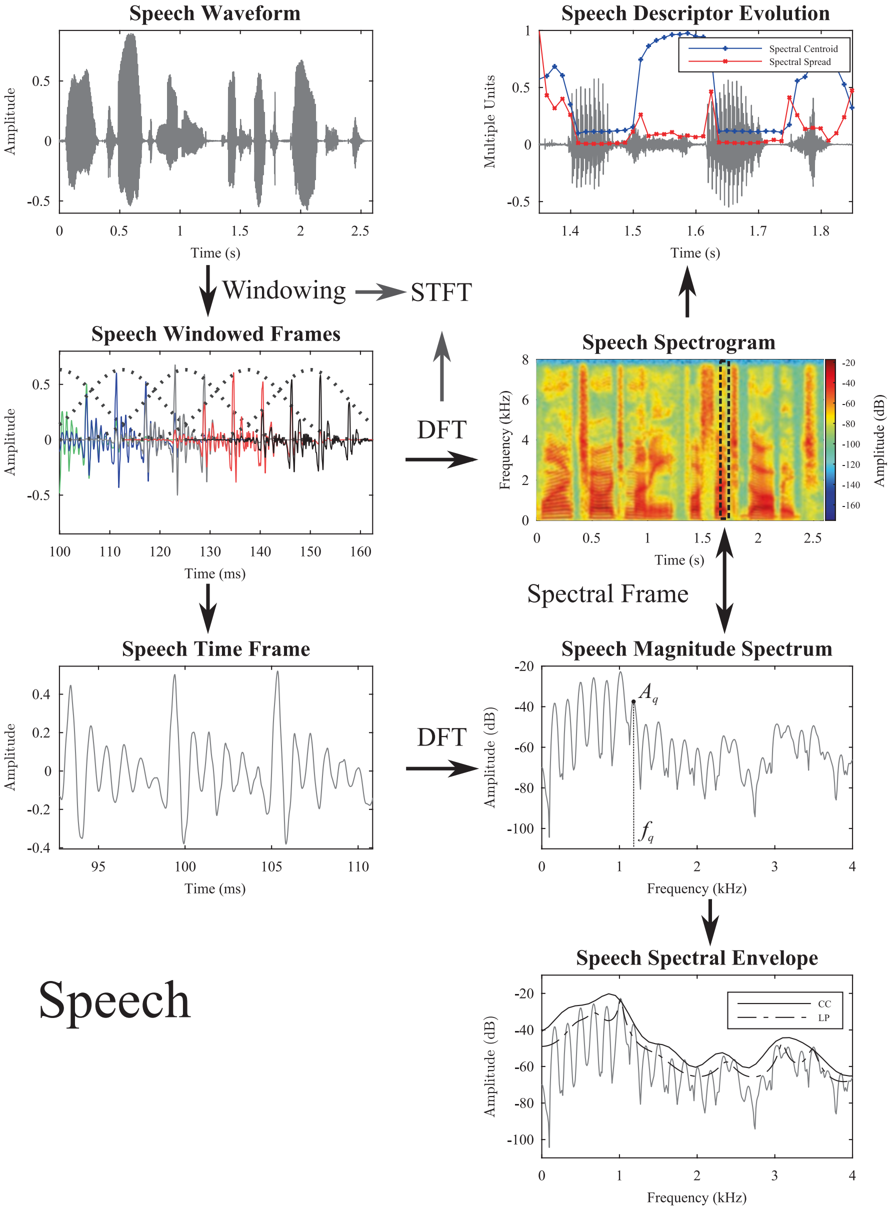

Fig. 11.2 Illustration of the sequence of steps to extract audio content descriptors from speech. The speech utterance consists of a 59-year-old white female speaker (U.S. southern accent) saying */why charge money for such garbage/*, sound file "speech.mp3". The extraction of descriptors from time-frequency representations is illustrated going *counter-clockwise* from the panel labeled *waveform* and the extraction of descriptors from the excitation-filter model is illustrated going *clockwise* from the panel labelled *time frame*. Abbreviations: A_q, amplitude of qth partial; *CC*, cepstral coefficients; *DFT*, discrete Fourier transform; f_q, frequency of qth partial; *LP*, linear prediction; *STFT*, short-time Fourier transform

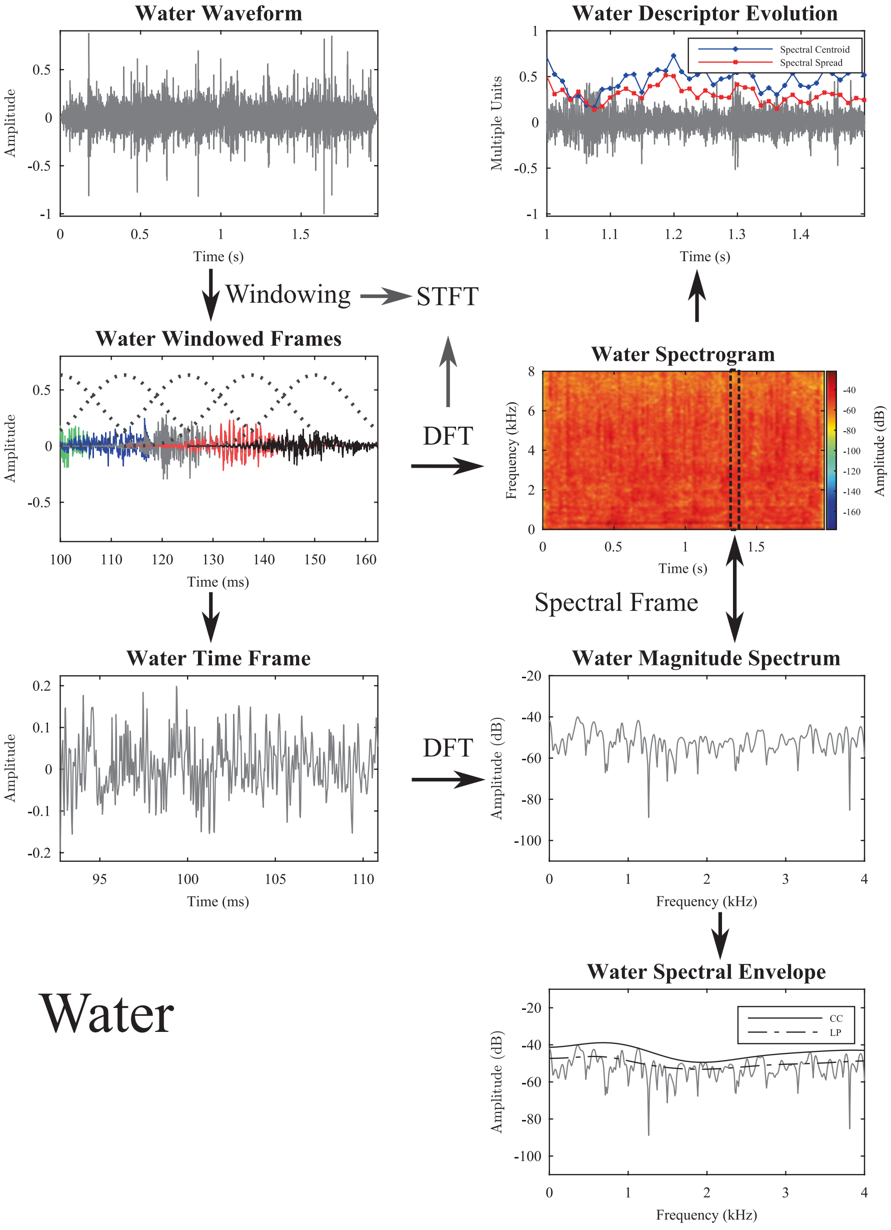

Fig. 11.3 Illustration of the sequence of steps to extract audio content descriptors from environmental sounds. The sample sound is running water, sound file "water.mp3". The extraction of descriptors from time-frequency representations is illustrated going *counter-clockwise* from the panel labeled *waveform* and the extraction of descriptors from the excitation-filter model is illustrated going *clockwise* from the panel labelled *time frame*. Abbreviations: A_q, amplitude of qth partial; *CC*, cepstral coefficients; *DFT*, discrete Fourier transform; f_q, frequency of qth partial; *LP*, linear prediction; *STFT*, short-time Fourier transform

Figs. 11.1–11.3 show the *spectrogram*, a visualization of the STFT in which the magnitude spectrum of each DFT is plotted against time, while amplitude information (in dB) is mapped to color intensity. The STFT has become the de facto analysis tool for various speech and music processing tasks. However, the STFT has notorious limitations for spectral analysis, mainly due to the constant length of the window.

The STFT is inherently limited by the *Fourier uncertainty principle*, a mathematical relation stating that a function and its Fourier transform cannot both be sharply localized (Jaffe 1987b). For audio processing, this implies that there is a fundamental tradeoff between time and frequency information. The constant length of the window in the STFT results in fixed temporal and spectral resolutions. Intuitively, frequency is a measure of the number of periods (or cycles) per unit time. Longer windows span more periods, which increases the accuracy in frequency estimation while simultaneously decreasing the temporal localization of the measurement. Therefore, time-frequency uncertainty is at the core of Fourier analysis and only a priori knowledge about the analyzed signal type and about the spectral properties of the window used (Harris 1978) can help choose the most appropriate spectral analysis tools for a specific application.

11.2.2 Constant Q Transform

The STFT can be interpreted as a filter bank with constant bandwidth and linear separation of the center frequency of each filter (Portnoff 1980; Dolson 1986). The constant bandwidth of each filter is a direct consequence of the fixed window length, whereas the linear separation of their center frequencies is due to the constant frequency bins of the DFT. However, the frequency intervals of Western musical scales are geometrically spaced (Brown 1991), so the frequency bins of the STFT do not coincide with the musical notes of Western musical scales. Additionally, the constant bandwidth of the STFT imposes a tradeoff in time-frequency resolution, where a window length naturally results in better spectral resolution for higher frequencies at the cost of poorer temporal resolution. In practice, each octave would require a different window length to guarantee that two adjacent notes in the musical scale that are played simultaneously can be resolved. The *constant Q transform* exploits a nonlinear frequency separation with an adaptive window length to yield a more compact representation of Western musical scales (Brown 1991). The quality factor of a resonator, denoted Q, is defined as the resonance frequency divided by the bandwidth of the resonator. The resonance frequency is the frequency at which the peak gain occurs, whereas the bandwidth is the frequency range around the resonance frequency where the gain is above a predefined threshold. The higher the Q, the narrower and sharper the peak is.

The constant Q transform can be calculated similarly to the DFT with geometrically spaced frequency bins and frames with lengths that depend on the analysis frequency. For musical applications, the frequency separation can be based on the musical scale with the semitone spacing of the equal tempered scale. A constant Q in the frequency domain corresponds to a frame length that is inversely proportional to frequency because the constant Q transform is designed to span the same number of periods inside each time frame. Thus, the constant Q transform is equivalent to a filter bank with adaptive bandwidths and nonlinear center frequencies in which the center frequencies can be aligned with the musical scale and the bandwidths are proportional to the center frequencies to yield a similar spectral resolution across all octaves.

Despite being useful for the spectral analysis of Western music, the original constant Q transform algorithm (Brown 1991; Brown and Puckette 1992) remained less popular than the STFT for two main reasons. Firstly, the constant Q transform was computationally inefficient compared to the fast Fourier transform (FFT) commonly used to calculate the STFT. Secondly, the original constant Q transform (Brown and Puckette 1992) was not invertible—it allowed sound analysis but not resynthesis. Recently, however, an efficient real-time implementation of a fully invertible constant Q transform was made possible using the concept of Gabor frames (Holighaus et al. 2013).

11.2.3 Auditory Filter Banks

The concepts of auditory filter banks and critical bands of human hearing are closely related to spectrum analysis over nonlinear frequency scales. Auditory filter banks model the acoustic response of the human ear with a bank of nonuniform bandpass filters whose bandwidths increase as the center frequency increases (Lyon 2017). Critical bands correspond to equal distances along the basilar membrane and represent the frequency bands into which the acoustic signal is split by the cochlea. Zwicker (1961) proposed the *Bark scale* to estimate the value of the first 24 critical bands as a function of center frequency based on empirical measurements using two-tone masking of narrowband noise. The Bark scale is approximately linear for frequencies below about 500 Hz and close to logarithmic at higher frequencies. Later, Glasberg and Moore (1990) suggested the *equivalent rectangular bandwidth* (ERB) scale for critical band estimation based on measurements using notched-noise masking. The ERB of a given auditory filter is defined as the bandwidth of a rectangular filter with similar height (peak gain) and area (total power) as the critical band it models. The ERB values are similar to those obtained by the Bark scale for center frequencies above 500 Hz, but they are markedly narrower at lower frequencies and thus more consistent with critical bandwidths measured with the more precise notched-noise method.

Gammatone filters are a popular choice to model the shape and frequency response of auditory filters because of their well-defined impulse response. A gammatone is a simple linear filter defined in the time domain as a waveform with an amplitude envelope having the shape of a gamma distribution. Patterson et al. (1992) showed that certain gammatone shapes provide a nearly perfect approximation to the measured human auditory filter shapes. A more precise approximation is obtained by *gammachirp filters*, in which the sinusoid is replaced by a monotonically frequency-modulated signal (i.e., a chirp) (Irino and Patterson 1997). Compared to the STFT, ERB-spaced gammatone filter banks offer a physiologically more accurate representation of the audio signal from which to extract spectral descriptors (Peeters et al. 2011; Siedenburg et al. 2016b). Nevertheless, the use of auditory filter banks in acoustic analysis for timbre remains less widespread than the STFT or cepstrum-based techniques, which are more straightforward to implement and are perfectly invertible.

11.2.4 Sinusoidal Modeling

Sinusoidal models (McAulay and Quatieri 1986) are a convenient representation of sounds that feature periodicity, such as musical instrument sounds and speech (see Figs. 11.1, 11.2) under the assumption that the sinusoids capture locally periodic oscillations in the waveform. In essence, sinusoidal models represent spectral peaks with sinusoids because the DFT of a sinusoid appears as a peak in the magnitude spectrum (Jaffe 1987b). The *time frame* panels show that musical instruments (Fig. 11.1) and speech (Fig. 11.2) feature relatively stable periodic oscillations (locally), whereas environmental sounds rarely do (Fig. 11.3). The amplitude and frequency of each spectral peak (see the *magnitude spectrum* panels in Figs. 11.1–11.3) are estimated for each frame (McAulay and Quatieri 1986). The partials are called harmonics when their frequencies are integer multiples of a fundamental frequency. The sum of all time-varying amplitudes of the partials gives the temporal envelope of the sound (see Fig. 11.4).

11.2.5 Temporal Envelope

The temporal amplitude envelope follows fluctuations of the amplitude of a signal. Mathematically, it is possible to express a signal as a combination of a slowly varying envelope and a rapidly varying carrier signal. The temporal envelope and time-varying phase of this representation of the signal are useful in audio descriptor extraction because they model amplitude and phase modulations, respectively (Elhilali, Chap. 12). Tremolo and vibrato are also intrinsically related to these parameters. For example, Regnier and Peeters (2009) proposed to use vibrato to automatically detect a singing voice in polyphonic music.

The *Hilbert transform* and the closely related analytic signal are useful tools to estimate the temporal envelope without prior sinusoidal modeling (Peeters et al. 2011). A fundamental property of the DFT is behind the connection between the original signal and the analytic signal derived from it. The DFT of a real signal is complex and its magnitude spectrum is symmetric around the frequency axis, as

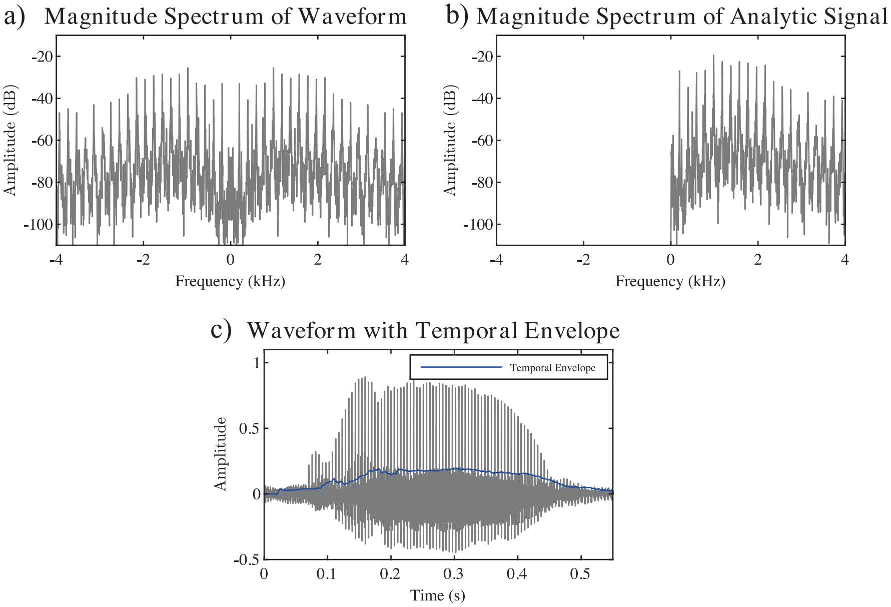

Fig. 11.4 Illustration of the analytic signal method to estimate the temporal amplitude envelope: (**a**) the magnitude spectrum of one of the trumpet notes from Fig. 11.1; (**b**) the magnitude spectrum of the analytic signal associated with (**a**); (**c**) the original waveform and the corresponding temporal envelope

shown in Fig. 11.4a. Mathematically, the property of symmetry means that the magnitude spectrum has a negative frequency component that has no physical interpretation. However, removing the negative frequencies and breaking the symmetry (see Fig. 11.4b) results in a spectrum that does not correspond to the original real signal anymore. In fact, the inverse DFT of the spectrum shown in Fig. 11.4b is a complex signal called the *analytic signal*, whose real part is the original signal and whose imaginary part is the Hilbert transform of the original signal. The temporal amplitude envelope can be calculated as the low-pass filtered magnitude of the analytic signal (Caetano et al. 2010; Peeters et al. 2011). Fig. 11.4c shows one of the trumpet notes seen in Fig. 11.1 with the temporal envelope calculated with the Hilbert transform. The Hilbert transform figures among the most widely used methods to estimate the temporal envelope, but it is hardly the only one (Caetano et al. 2010).

11.2.6 Excitation-Filter Model and Convolution

The excitation-filter model, also called the source-filter model (Slawson 1985; Handel 1995), offers a simple yet compelling account of sound production whereby a driving force, the excitation, causes a physical object, the filter, to vibrate. Here the term *excitation* is preferred over *source* to avoid potential confusion with the source of a sound, such as a musical instrument or a person. The physical properties of the vibrating object cause it to respond differently to different frequencies present in the excitation. Consequently, the vibrating object acts as a filter on the excitation, attenuating certain frequencies while emphasizing others. For example, a knock on a door is a short abrupt driving force that causes the door to vibrate. The sound from a wooden door is different from the sound from one with similar dimensions made of glass or metal due to their different material properties. Bowing the strings of a violin will cause its body to vibrate; the shape, size, and material of the violin body are responsible for the unique sonority of the instrument. Similarly, air through the lungs causes the vocal folds to vibrate, and the vocal tract shapes these vibrations into the unique timbre of a person's voice.

The interaction between the properties of the excitation and those of the vibrating object can be interpreted as filtering, which is mathematically expressed as a *convolution*. The Fourier transform is the key to understanding why convolution is mathematically equivalent to filtering because convolution in the time domain becomes multiplication in the frequency domain (Jaffe 1987b). This property of convolution is extremely useful for the analysis of sounds and the extraction of audio content descriptors of timbre in light of the excitation-filter model. In particular, the filter component (or transfer function) models how the physical properties of the vibrating object respond to the excitation in the frequency domain. The contributions of the excitation and filter can theoretically be isolated in the frequency domain and inverted, bringing the frequency spectrum back to the time domain. In the time domain, the transfer function is called the *impulse response* and is essentially a model of the physical properties of the vibrating object. Consequently, the impulse response carries information intrinsically related to timbre perception that can be used to extract audio descriptors of timbre. Section 11.3.3 explores some of the most widely used audio descriptors of timbre based on the excitation-filter model.

11.3 Extraction of Timbre Descriptors

The raw information provided by audio signal representations such as the STFT and the excitation-filter model is usually not specific enough to describe salient aspects of timbre. Therefore, a plethora of techniques for extracting timbre-relevant descriptors from these representations has been proposed in the field of audio content analysis. Some audio descriptors are extracted from generic time-frequency representations and are later found to capture aspects of timbre perception, whereas others are based on the excitation-filter model and commonly describe physical properties of the sound source. In general, audio descriptors can represent global or local aspects of sounds. *Global descriptors* only have one value for the entire duration of a sound, whereas *local descriptors* are commonly calculated for every frame (see Figs. 11.1–11.3) and result in a time series.

Additionally, descriptors can be categorized as temporal, spectral, or spectrotemporal (Peeters et al. 2011). *Temporal descriptors* exclusively capture temporal aspects of sounds and are generally global. Some are computed directly from the waveform, but most are typically extracted from the temporal energy envelope (Sect. 11.2.5). *Spectral descriptors* capture local features of the frequency content regardless of the surrounding frames. Spectral descriptors also have an alternative harmonic version calculated from the sinusoidal model. Finally, *spectrotemporal descriptors* capture spectral changes relative to the previous or next frames. Thus, spectrotemporal descriptors attempt to incorporate time as relative local spectral changes throughout the duration of a sound.

Section 11.3.1 addresses temporal descriptors and Sect. 11.3.2 covers descriptors extracted from time-frequency representations. Section 11.3.3 focuses on descriptors based on the excitation-filter model, Sect. 11.3.4 explores the temporal dynamics of the time series of descriptors, and Sect. 11.3.5 discusses information redundancy among descriptors.

11.3.1 Temporal Descriptors

The zero-crossing rate is a measure of how many times the waveform changes sign (i.e., crosses the zero axis). In general, periodic sounds have a smaller zero-crossing rate than noisier sounds, so the zero-crossing rate can be used in voice activity detection, voiced-unvoiced decisions for speech, and even in the classification of percussive sounds (Peeters et al. 2011), although there is no straightforward perceptual interpretation of the zero-crossing rate (Siedenburg et al. 2016a).

The temporal envelope is used to extract temporal descriptors such as tremolo (Peeters et al. 2011), the temporal centroid, and attack time. The *temporal centroid* is the temporal counterpart of the spectral centroid (see Sect. 11.4.2). Percussive sounds have a lower temporal centroid than sustained sounds. McAdams et al. (2017) found that a lower (i.e., earlier) temporal centroid correlated strongly with the valence of musical affect carried by the timbre of musical instrument sounds.

The *attack time* is the time between the onset of a sound and its more stable part. In musical instruments, for example, the attack time accounts for the time the partials take to stabilize into nearly periodic oscillations. Percussive musical instruments, such as the xylophone, feature short attack times with sharp onsets, whereas sustained instruments, such as the tuba, feature longer attack times. The attack time of a waveform can be estimated with models such as the weakest effort method (Peeters et al. 2011) or the amplitude/centroid trajectory model (Hajda 2007; Caetano et al. 2010). The *weakest effort method* uses signal-adaptive energy thresholds instead of fixed energy levels to estimate the beginning and end of the attack from the temporal envelope. The *amplitude/centroid trajectory model* uses spectrotemporal information from both the temporal envelope and the temporal evolution of the spectral centroid to segment musical instrument sounds. Section 11.3.3 will delve deeper into the amplitude/centroid trajectory model and the timbre descriptors used therein. The attack time consistently arises as one of the most salient dimensions in timbre spaces from MDS studies (Grey 1977; Siedenburg et al. 2016a).

McAdams et al. (1995) found the logarithm of the attack time among the most salient dimensions of perceptual dissimilarity.

Other common temporal descriptors (Peeters et al. 2011) include the slopes of the energy envelope during the attack and decrease segments, the effective duration, and the temporal modulation of energy over time (i.e., tremolo). Energy modulation is calculated either from the temporal evolution of the amplitudes of isolated partials across frames or from the temporal envelope.

11.3.2 Time-Frequency Representations and Audio Descriptors

11.3.2.1 Spectral Descriptors

Spectral descriptors are typically calculated for each frame of a time-frequency representation such as the STFT (see Figs. 11.1–11.3). Descriptors of spectral shape characterize the overall spectral distribution of sounds and are calculated as if the STFT magnitude spectrum were a probability distribution. Peeters et al. (2011) remarked that spectral descriptors can use different spectral scales such as magnitude, power, or log.

The spectral shape descriptors, calculated similarly to the standardized moments of the frequency spectrum, are the spectral centroid, spectral spread, spectral skewness, and spectral kurtosis. The *spectral centroid* is the amplitude-weighted mean frequency. It is measured in Hertz (Hz) and is analogous to the center of mass, so it can be interpreted as the center of balance of spectral energy distribution or the frequency that divides the spectrum into two regions with equal energy. The spectral centroid often appears among the most salient dimensions of timbre spaces (see McAdams, Chap. 2), and it is interpreted as capturing the “brightness” of a sound (Grey and Gordon 1978; McAdams et al. 1995). Sounds described as bright, such as a brassy trombone note, have higher spectral centroids because they feature more spectral energy in high frequency regions (see Sect. 11.4.1). *The spectral spread* measures the spread of spectral energy around the spectral centroid. It is related to the bandwidth of a filter, so a brighter sound will have a higher spectral spread than a duller sound. *Spectral skewness* is a measure of asymmetry of spectral energy around the spectral centroid. Negative values indicate more spectral energy concentrated at frequencies lower than the spectral centroid, positive values indicate more energy at higher frequencies than the centroid, and zero indicates energy symmetry around the centroid. Finally, *spectral kurtosis* is a measure of the flatness of the spectral distribution of energy compared to a normal distribution. A negative value of spectral kurtosis indicates a distribution of spectral energy flatter than the normal distribution, whereas a positive value indicates the opposite.

Spectral flux or *spectral variation* (Casey et al. 2008; Peeters et al. 2011) is considered a spectrotemporal descriptor because it captures local spectral change over time. Essentially, it measures the spectral difference of the current frame relative to the previous frame. Compared to attack time and spectral centroid, the correlation of spectral flux with listener ratings of timbre similarity has been less consistent. While in some studies the third dimension of the MDS timbre space does correlate moderately well with the time-varying spectral flux (McAdams et al. 1995), in others it correlates better with static descriptors of *spectral deviation* (deviation of partial amplitudes from a global, smoothed spectral envelope; Krimphoff et al. 1994) or *spectral irregularity* (attenuation of even harmonics; Caclin et al. 2005).

Several other spectral descriptors appear in the literature (Peeters et al. 2011), many of which capture similar properties of the spectrum. However, there is little consensus about their usefulness or even their relationship with timbre (see Sect. 11.4.1).

11.3.2.2 Harmonic Content

Most spectral descriptors also have a harmonic version calculated by simply replacing the spectral magnitude with the amplitudes of the sinusoidal model (see the *magnitude spectrum* panels in Figs. 11.1–11.3), such as the harmonic energy (Peeters et al. 2011). However, some descriptors capture information specifically related to the oscillatory modes of the signal, commonly called *partials*. Figs. 11.1–11.3 highlight the differences in both time and frequency domains for sounds of musical instruments, speech, and environmental sounds (represented by running water in Fig. 11.3). The *time frame* panels reveal that both musical instruments and speech feature relatively stable oscillations in some regions (except where changes, such as note transitions, are happening), whereas the running water sound is noisy. Oscillations in the time domain appear as spectral peaks in the frequency domain. The magnitude spectrum of the musical instrument shows prominent spectral peaks across the entire frequency range of 0–4 kHz. For speech, the spectral peaks are less prominent beyond approximately 2.2 kHz. Finally, the magnitude spectrum for the water sound shows a relatively flat distribution of spectral energy typical of noise.

A fundamental result from Fourier analysis (Jaffe 1987a) reveals that the spectrum of a perfectly periodic waveform is perfectly harmonic. However, neither speech nor musical instrument sounds are perfectly periodic. Consequently, neither type has a spectrum that features perfectly harmonic spectral peaks. This can be quantified with the descriptor *inharmonicity*, based on the sinusoidal model (see Sect. 11.2.4). Inharmonicity measures the deviation of the frequencies of the partials from pure harmonics, calculated as the normalized sum of the differences weighted by the amplitudes (Peeters et al. 2011). Sustained musical instruments, such as those from the woodwind (e.g., flute, clarinet, bassoon, and oboe), brass (e.g., trumpet, trombone, and tuba), and string (e.g., violin, viola, and cello) families, produce sounds whose spectra are nearly harmonic (Fletcher 1999). Percussion instruments (e.g., cymbals and timpani) are considered inharmonic, whereas others (e.g., bells or the piano) feature different degrees of inharmonicity (Fletcher 1999; Rigaud & David 2013). The spectrum of the piano, for example, has partials whose inharmonicity is proportional to the partial number. So, the higher the frequency, the greater is the deviation from the harmonic series (Rigaud & David 2013). This characteristic inharmonicity is an essential property of the timbre of the piano. Features that are specific to certain musical instruments, commonly called *specificities*, are directly related to timbre perceptions of these instruments. For instance, the timbre of clarinet sounds, described as “hollow” (McAdams et al. 1995), can be linked to spectral energy predominantly concentrated around odd harmonics. The *odd-to-even harmonic energy ratio* (Peeters et al. 2011) is a descriptor that quantifies this particular specificity.

Pollard and Jansson (1982) proposed a three-dimensional representation of timbre dubbed *tristimulus*. Each dimension of the tristimulus representation contains the loudness of a group of partials (i.e., how much energy each group contributes to the overall spectrum). The first dimension has the fundamental frequency, the second dimension includes partials two to four, and the third dimension contains the rest of the partials from the fifth to the highest. Pollard and Jansson (1982) used the tristimulus method to represent the temporal evolution of musical instrument sounds and revealed variations in timbre with time, especially between the attack transients and the steady state with its more stable oscillatory behavior. Section 11.3.4 will explore further the temporal evolution of descriptors and timbre.

11.3.3 The Excitation-Filter Model and Audio Descriptors

There are several descriptors of timbre based on the excitation-filter model of sound production (introduced in Sect. 11.2.6). These descriptors typically capture information related to the filter component of the model, which is responsible for the relative distribution of spectral energy. Perceptually, the relative energy of spectral components is directly related to timbre and is sometimes called sound color (Slawson 1985). When associated with the excitation-filter model, the spectral envelope (see the *spectral envelope* panels in Figs. 11.1–11.3) is commonly used to represent the filter component.

Descriptors of timbre based on the excitation-filter model commonly use the magnitude spectrum and discard the phase, autocorrelation coefficients being the quintessential example. Autocorrelation is a measure of self-similarity, whereby a signal is compared with its own past and future values. The autocorrelation and convolution operations share similarities that become more evident with the DFT (Jaffe 1987b). The autocorrelation coefficients are the time domain representation of the power spectral density (Jaffe 1987b; Brown et al. 2001), so they are related to the filter component. The relationship between autocorrelation coefficients and power spectral density is exploited further by linear prediction (Makhoul 1975).

11.3.3.1 Linear Prediction Coefficients

Linear prediction (Makhoul 1975) assumes that a signal can be described as a weighted linear combination of past values plus an external influence. The external influence accounts for the force exciting the vibrating object that generated the signal, whereas the vibrating object itself is not explicitly modeled. When the external influence is unknown, the signal can only be approximated by its past values. The model parameters can be estimated by minimizing the mean squared error. The solution yields the set of *linear prediction coefficients* (LPC) that best predict the next value of the signal given a specific number of preceding values in the least squared error sense, which is mathematically equivalent to using the autocorrelations to estimate the LPC (Makhoul 1975).

The LPC are commonly represented in the frequency domain with the Z-transform, which encodes essentially the same information as the Fourier transform (Jaffe 1987b) but in a more general framework. Similarly to the DFT, the Z-transform can also be interpreted as the frequency domain representation of the signal. The Z-transform of the linear prediction model explicitly reveals the frequency response of the vibrating object under the force that resulted in the sound spectrum. This frequency response is the filter component, commonly called the transfer function, and it fully characterizes the model of sound production under certain assumptions (Makhoul 1975).

Immediate physical and physiological interpretations for musical instrument sounds and speech can be derived from the LPC. For example, the LPC can be interpreted as a model of the resonances of the vocal tract in speech production (Makhoul 1975) because they encode the poles of the filter that approximates the original power spectral density. Linear prediction is commonly used to approximate the power spectrum with the spectral envelope (see the LP curve in the *spectral envelope* panels in Figs. 11.1–11.3), defined as a smooth curve that approximately connects the spectral peaks (Burred et al. 2010).

11.3.3.2 The Cepstrum

The cepstrum (Bogert et al. 1963; Childers et al. 1977) is intimately connected with the excitation-filter model because it was originally developed as a deconvolution method (Bogert et al. 1963). The excitation-filter model postulates that a waveform can be described mathematically as the convolution between the filter and the excitation. Deconvolution allows recovery of either the filter or the excitation from the waveform. In the frequency domain, convolution becomes multiplication (see Sect. 11.2.6) and deconvolution becomes inverting the result of the multiplication. Division is the simplest method when the DFT of either the excitation or the filter is available, allowing recovery of the other. However, in most practical applications, only the waveform resulting from the convolution between the excitation and the filter is available. In this case, the logarithm can be used to transform the multiplication operation into addition. If the terms of the resulting addition do not overlap in frequency, it is possible to isolate either one from the other by filtering. The *cepstrum* is the formalization of this deconvolution operation (Childers et al. 1977), which has found several applications in audio research, such as fundamental frequency estimation (Childers et al. 1977), spectral envelope estimation (Burred et al. 2010), wavelet recovery (Bogert et al. 1963), and musical instrument classification (Brown 1999; Herrera-Boyer et al. 2003). The *spectral envelope* panels in Figs. 11.1–11.3 show its estimation with the cepstrum (identified as *CC*).

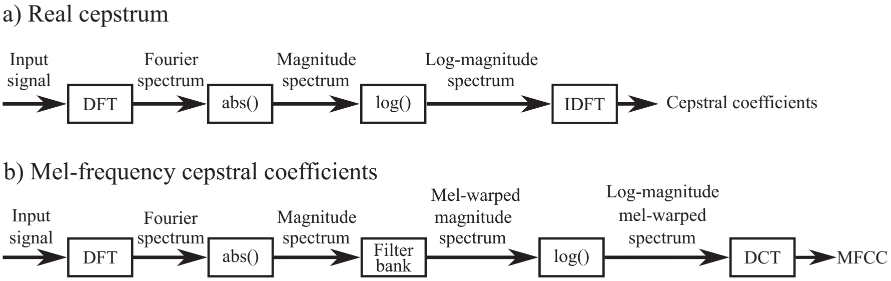

Fig. 11.5 Illustration of the sequence of steps to calculate cepstral coefficients with the real cepstrum (**a**) and mel-frequency cepstral coefficients (MFCC) (**b**) from the waveform. Abbreviations: *abs()*, absolute value; *DCT*, discrete cosine transform; *DFT*, discrete Fourier transform; *IDFT*, inverse discrete Fourier transform; *log()*, logarithm function; *MFCC*, mel-frequency cepstral coefficients

The real cepstrum can be defined as the inverse DFT of the logarithm of the magnitude of the DFT of the waveform. Fig. 11.5 illustrates the steps to obtain cepstral coefficients from a waveform (labeled *input signal*). The cepstral coefficients contain frequency information about the log magnitude spectrum similarly to how the LPC encode the resonances of a transfer function. In practice, these coefficients encode information about periodicity of the log magnitude spectrum at increasing cepstral frequencies, which were originally called “quefrencies” (Bogert et al. 1963), because they carry frequency information in time domain units. This unfamiliar symmetry was reflected in language by rearranging syllables of familiar terms from Fourier analysis. Particularly, “cepstrum” derives from spectrum and is pronounced *kepstrum*.

11.3.3.3 Mel-Frequency Cepstral Coefficients

Conceptually, the cepstral coefficients are closely related to the filter component of the excitation-filter model and of the ubiquitous mel-frequency cepstral coefficients (MFCC; mel is short for melody). Davies and Mermelstein (1980) introduced MFCC in the context of speech research. The MFCC can be viewed as a perceptually inspired variation of cepstral coefficients calculated as illustrated in Fig. 11.5. The MFCC filter bank uses triangular filters centered at frequencies given by the mel scale with a bandwidth proportional to the center frequency.

The perception of pitch allows listeners to order sounds on a scale from low to high along the same psychological dimension of melody (Hartmann 1996). A sound has a certain pitch if it can be reliably matched by adjusting the frequency of a sine wave of arbitrary amplitude (Hartmann 1996). The mel scale was derived by asking listeners to set the frequency of a test sine wave to obtain a pitch that was a fraction of the pitch of a reference sine wave across the entire audible frequency range (approximately between 20 Hz and 20 kHz). It is linear up to 1 kHz and logarithmic above 1 kHz. Stevens et al. (1937) concluded that the mel scale captures the concept of pitch height (i.e., higher or lower pitches) as opposed to pitch chroma (i.e., the octave-independent musical notes). The MFCC use the discrete cosine transform (commonly used in MPEG audio and image compression) instead of the DFT or the Z-transform commonly used for cepstral coefficients. Thus, MFCC are considered a particularly compact representation of the filter due to the compression properties of the discrete cosine transform, which results in most of the spectral shape being captured typically by the first thirteen coefficients.

The MFCC are ubiquitous not only in speech for tasks such as speaker recognition (On et al. 2006; Martínez et al. 2012) but also in MIR tasks such as musical instrument classification (Deng et al. 2008). Some results suggest that MFCC can also explain timbre. For example, Terasawa et al. (2005) compared MFCC, LPC, and tristimulus (see Sect. 11.3.2.2) representations to explain the pairwise perceptual dissimilarity ratings of sounds created with frequency-modulation synthesis. They found that the Euclidean distances between MFCC accounted for 66% of the variance and concluded that thirteen MFCC can be used as a model of timbre spaces. Horner et al. (2011) compared different error metrics to predict the discrimination performance of listeners for sounds synthesized with fixed fundamental frequency and variable spectral envelope. They found that the first twelve MFCC were sufficient to account for around 85% of the variance of data from human listeners.

11.3.4 Temporal Dynamics of Audio Descriptors

Many descriptors are calculated for every frame of time-frequency representations, such as the STFT, giving rise to a time series of descriptor values that characterizes the temporal evolution of each descriptor. The *descriptor evolution* panels in Figs. 11.1–11.3 show the temporal evolution of the spectral centroid and spectral spread, revealing local variations corresponding to changes such as note transitions. However, most applications, such as musical instrument classification, require one single value of each descriptor that would be representative of the entire sound duration. Commonly, the time average of each descriptor is used for each sound, resulting in a descriptor vector. Descriptors such as the spectral centroid are unidimensional, whereas others, such as MFCC, are multidimensional. Therefore, descriptor vectors discard all information about the temporal variation of descriptors.

The simplest way to include more information than the time average of the descriptors is to use a set of summary statistics such as mean, standard deviation (or variance), minimum, and maximum values (Casey et al. 2008). Peeters et al. (2011) found that robust summary statistics had a greater impact than the audio representation on the descriptors. Specifically, the median and the interquartile range captured distinct aspects of the signals. McDermott et al. (2013) suggested that environmental sounds are recognized by summary statistics alone because the temporal information in environmental sounds can be captured by summary statistics. However, the temporal structure inherent to speech and musical sounds requires encoding temporal information in different ways.

The first and second derivatives with respect to time (of the time series of descriptor values) are another popular approach to include temporal information in the descriptor vector. It is particularly common to use MFCC and their first and second temporal derivatives, called delta and delta-delta coefficients, respectively (De Poli and Prandoni 1997; Peeters et al. 2011). However, the delta and delta-delta coefficients are usually added to the descriptor vector as extra dimensions assumed to be independent from the descriptor values. Consequently, the information contained in the time series of descriptor values is not fully exploited. For example, Fig. 11.2 reveals that the spectral centroid of speech varies considerably between periodic and noisier segments. Similarly, for musical instruments, the temporal variation of descriptors follows musical events such as note transitions. The amplitude/centroid trajectory model (Hajda 2007) shown in Fig. 11.6 proposes to use the root-mean-squared amplitude envelope in conjunction with the temporal evolution of the spectral centroid to segment sustained sounds from musical instruments into attack, transition (so-called decay), sustain, and release portions. Fig. 11.6 shows the amplitude-centroid trajectory model used to segment notes from sustained musical instruments (Caetano et al. 2010). Segmentation of musical instrument sounds with the amplitude-centroid trajectory model yields better results for sustained instruments than percussive ones because sustained instruments fit the model better.

The use of a descriptor vector with the time average of each descriptor in each dimension is called the *bag of frames approach* because it treats the time series of descriptors as a global distribution, neglecting both the temporal variation and the sequential order of descriptor values (Levy and Sandler 2009; Huq et al. 2010). This approach can be successfully used to classify environmental sounds (Aucouturier

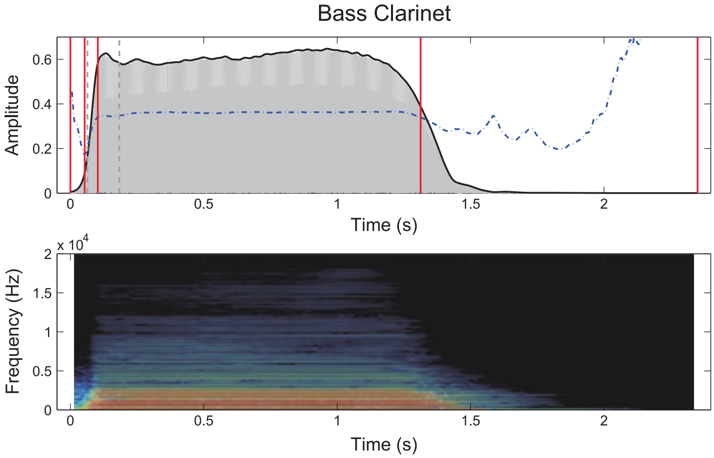

Fig. 11.6 Temporal segmentation of musical instrument sound with the Amplitude/Centroid Trajectory (ACT) model. Top panel: the full-wave rectified waveform outlined by the temporal amplitude envelope (*solid line*) and the temporal evolution of the spectral centroid (*dashed line*). The vertical bars mark the segments estimated with the ACT method. See text for an explanation of the segments. Bottom panel: the spectrogram of the waveform on the top. (Reproduced from Caetano et al. 2010; used with permission)

et al. 2007) with Gaussian mixture models representing the global distribution of MFCC. However, it is inappropriate for polyphonic music (Aucouturier et al. 2007) in which the sequence of events contains important information. In music, there is a clear hierarchical structure where higher levels of abstraction emerge from lower levels. For example, patterns of notes are organized into phrases, and rhythmic structure emerges from relative note durations. Aucouturier et al. (2007) speculated that the hierarchical structure of polyphonic music carries information on a more symbolic level than is captured by descriptors such as MFCC, requiring incorporation of information such as harmony and melody.

Temporal modeling of descriptors has been successfully applied in instrument classification and detection. Models of musical instrument sounds that rely on spectrotemporal representations are capable of capturing the dynamic behavior of the spectral envelope (Burred and Röbel 2010; Burred et al. 2010). Principal component analysis reduces the dimensionality of the model by projecting the time-varying parameters of the spectral envelopes onto a lower-dimensional space, such as the three-dimensional space shown in Fig. 11.7. The resultant prototypical temporal evolution of the spectral envelopes was modeled as a nonstationary Gaussian process and was shown to outperform MFCC for the classification of isolated musical instruments and to allow for instrument recognition in polyphonic timbral mixtures.

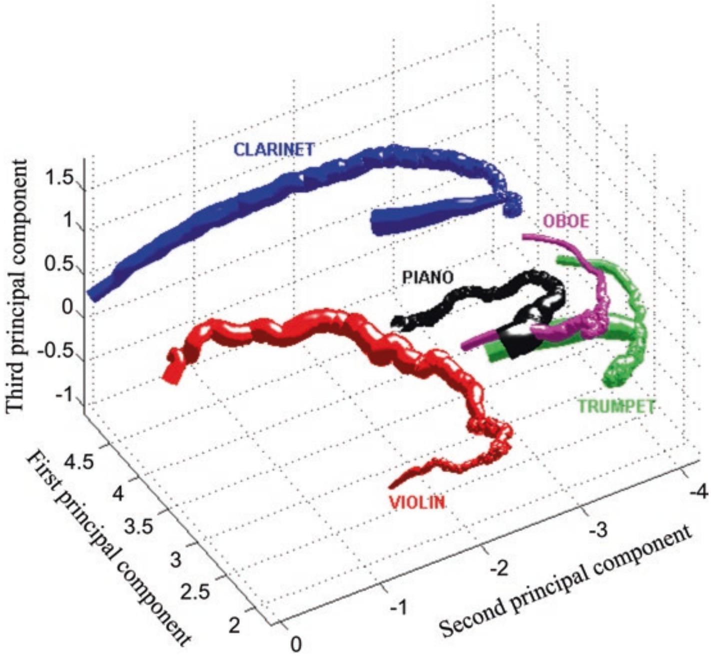

Fig. 11.7 Temporal evolution of the spectral envelope of musical instrument sounds. The temporal trajectory of the spectral envelope of the musical instruments indicated (*clarinet*, *oboe*, *piano*, *trumpet*, and *violin*) is shown in a three-dimensional representation obtained with principal component analysis. (Reproduced from Burred and Röbel 2010; used with permission)

11.3.5 Information Redundancy of Audio Descriptors

Descriptor vectors stack several descriptors under the assumption that each dimension is statistically independent from the others. While this assumption might hold true for some descriptors, such as MFCC, which are decorrelated by construction due to the discrete cosine transform (see Sect. 11.3.2.2), other descriptors are highly correlated. Peeters et al. (2011) investigated the correlation structure among descriptors extracted with alternative representations based on an analysis of over 6000 musical instrument sounds with different pitches, dynamics, articulations, and playing techniques. The authors observed that a change in the audio representation (e.g., STFT versus ERB-spaced filterbank versus harmonic content) had relatively little effect on the interdescriptor correlation compared to the change in the summary statistic computed on the time-varying descriptors, although no prediction of perceptual data was undertaken in that paper.

Several strategies have been proposed to decrease information redundancy in descriptor vectors. Among these, the most common ones fall generally into *descriptor selection* or *descriptor decomposition* strategies. Descriptor selection involves finding the subset of descriptors that is useful to build a good predictor (Huq et al. 2010) by eliminating descriptors that are either irrelevant or redundant. On the other hand, descriptor decomposition techniques apply transformations on the original space of descriptors that aim to maximize the information that is relevant for a task in the reduced space, such as the variance of the descriptors or the discriminability of classes. These transformations commonly involve projection or compression techniques, such as principal component analysis for the former and the discrete cosine transform for the latter. Descriptor decomposition techniques commonly distort the original representation in ways that can render interpretation more difficult. For example, principal component analysis results in linear combinations of the original dimensions that, in practice, render a perceptual interpretation of the results that is more arduous because each principal component accounts for more than one descriptor. Descriptor selection preserves the original meaning of the variables by preserving their original representation, ultimately offering the advantage of interpretability. At the same time, descriptor selection can lead to choices that seem arbitrary in that the selected descriptors may vary a great deal from one study to another.

11.4 Applications of Timbre Descriptors

Audio content descriptors find several applications that involve timbre description. Examples about the study of timbre psychoacoustics are discussed in Sect. 11.4.1; the multimedia content description interface also known as MPEG-7 is discussed in Sect. 11.4.2; computer-aided orchestration is discussed in Sect. 11.4.3; and musical instrument sound morphing guided by descriptors of timbre is discussed in Sect. 11.4.4.

11.4.1 Timbre Psychoacoustics

Audio signal descriptors have been central to the psychoacoustics of timbre, which seeks an explanation of timbre perception on acoustic grounds. Most of this research has used musical instrument sounds. A notable exception is the work by Zwicker and Fastl (1990), who presented acoustic models of sharpness, fluctuation strength, and roughness, which have been used mainly to characterize the sound quality of product sounds (see Lemaitre and Susini, Chap. 9). In the following discussion, three examples of using audio descriptors for psychoacoustics research will be addressed. These examples highlight the search for acoustic correlates of timbral brightness judgements and sound source recognition.

For musical sounds, two methods to study brightness perception can be distinguished. First, timbre space dimensions obtained via MDS of general dissimilarity judgements have consistently been interpreted as associated with the brightness of sounds (see McAdams, Chap. 2). Second, several studies have directly asked participants to rate the brightness of sounds and have correlated the resulting ratings with descriptor values. For instance, Schubert and Wolfe (2006) considered whether direct brightness ratings are better predicted by the absolute spectral centroid or the (supposedly pitch invariant) centroid rank (the centroid divided by the fundamental frequency). The latter predictor, however, failed to correlate significantly with subjective brightness, whereas the absolute centroid did.

Marozeau and de Cheveigné (2007) proposed a refined spectral centroid descriptor to model the brightness dimension of dissimilarity ratings. The approach was conceptually related to the sharpness descriptor by Zwicker and Fastl (1990) in that it relied on the computation of partial loudness in spectral bands (but the Zwicker model only insufficiently predicted brightness scaling data in Almeida et al. 2017). Specifically, the descriptor by Marozeau and de Cheveigné (2007) was obtained from partial loudness values calculated in ERB-spaced spectral bands obtained from gammatone filtering (see Sect. 11.2.3). An instantaneous spectral centroid was obtained through the integration across bands and the resulting time series was integrated over time by weighting with an estimate of instantaneous loudness (the sum over channels of partial loudness). In comparison to the linear spectral centroid descriptor, the refined brightness descriptor by Marozeau and de Cheveigné (2007) improved the amount of the explained variance with the perceptual data by 10% points up to 93%. Further analysis showed that it was much less affected by pitch variation compared to the more simplistic linear spectral centroid.

Fewer studies have used signal descriptors to address the acoustic features underlying sound source recognition and classification (see Agus, Suied, and Pressnitzer, Chap. 3). Ogg et al. (2017) modeled participant responses in a go/no-go categorization task of short sound excerpts varying in duration (12.5–200 ms). Three sound categories were tested: speech, musical instruments, and human environmental sounds generated by everyday objects (e.g., keys jingling), by objects of various materials impacting one another or being deformed (e.g., crumpling newspaper), and sounds generated by movements of liquid (fingers splashing) or rolling objects (marbles rolling down wood). Using exploratory regression analysis with timbre descriptors from the Timbre Toolbox (Peeters et al. 2011), the study characterized the acoustic features that listeners were using to correctly classify sounds. Globally, regression analysis for sounds from every target category indicated that listeners relied on cues derived from spectral, temporal, pitch, and “noisiness” information. Different sound categories required different sets of descriptors and weightings of regression coefficients. For instance, as the median spectral centroid value increased, listeners were more likely to categorize the stimuli as human environmental sounds and less likely to consider the sounds as coming from musical instruments. The descriptors “noisiness” and “spectral flatness” were associated with environmental and instrument responses, respectively.

Approaches such as these provide valuable starting points to reveal the most important acoustic features for a given psychophysical task from the plethora of available audio content descriptors. Giordano et al. (2012) further showed that audio descriptors can be applied to neuroimaging research (for neurophysiological details, see Alluri and Kadiri, Chap. 6). Following the approach of representational similarity analyses, they used descriptors to decode fMRI data recorded while participants listened to environmental sounds. They extracted descriptors based on pitch, loudness, spectral centroid, and harmonicity, and they computed dissimilarity matrices that contained the pairwise dissimilarity of stimuli according to these descriptors. Dissimilarity matrices were also derived from the imaging data, specifically, from the response of each voxel in a region of interest. Then, correlation of the neurophysiological and the acoustic dissimilarity matrices resulted in maps that indicated the association of the activity in a given voxel to a specific acoustic property. Hence, this approach can infer the brain areas associated with the processing of low-level acoustic properties represented by the audio descriptors.

These examples indicate that a variety of psychophysical and even psychophysiological questions on timbre can benefit from a deeper involvement with audio descriptors, which can be easily computed today (Peeters et al. 2011). At the same time, the correlational nature of the approach warrants rigorous confirmatory studies to circumvent the strong mutual covariance of descriptors.

More generally, it seems important to acknowledge that work on timbre-related audio content descriptors is at the crossroads of distinct academic fields, including MIR, music cognition, and psychoacoustics. Hence, it is important to appreciate the distinct epistemic traditions and objectives that are ingrained in these fields (Siedenburg et al. 2016a). Music information retrieval is a task-oriented discipline rooted in applied computer science and machine learning and, therefore, is primarily interested in the question of how to build robust systems. This implies that the predictive power of a descriptor is more important than the exact acoustic properties it encodes. In psychology, however, researchers are interested in the insights an audio descriptor can bring to the study of a given perceptual phenomenon. If a descriptor does not add significantly to the overall explanatory power of a model, and if the information it encodes is not transparent, then it should be omitted for the sake of parsimony. These considerations reflect some of the epistemic undercurrents of this topic and explain why studies on timbre psychoacoustics have traditionally used relatively fewer audio descriptors, whereas MIR research on automatic instrument classification used the full gamut of available descriptors.

11.4.2 MPEG-7 Audio Content Description

The multimedia content description interface (Nack and Lindsay 1999; Martínez et al. 2002), also known as MPEG-7, is part of a large effort to standardize multimedia descriptors and descriptor schemes that allow indexing and searching multimedia content, such as pictures, video, audio, and information about how those elements combine in a multimedia context. Unlike the previous MPEG standards that addressed coded representations of audiovisual information, MPEG-7 addresses the representation of information *about* the content, but not the content itself. MPEG-7 began as a scheme for making audiovisual material as searchable as text is today (Nack and Lindsay 1999) and grew to include complex scenarios that employ image processing (such as surveillance) and media conversion, for example, speech to text (Martínez et al. 2002). Within the audio domain, MPEG-7 provides a unified interface for automatic organization of audio from different multimedia sources (i.e., music and film) for applications in sound archiving and classification, and for retrieval, such as music indexing, similarity matching, and MIR (Casey 2001a, b). In addition to traditional timbre description methods that have been applied mainly to isolated musical instrument notes, MPEG-7 also represents noise textures, environmental sounds, music recordings, melodic sequences, vocal utterances (singing and speech), and audio mixtures of the above (Casey 2001b).

MPEG-7 audio comprises text-based description by category labels, also called semantic tags, and quantitative description using audio content descriptors, as explained in Sect. 11.3. Text-based description consists of semantic tags from human annotations (Casey 2001a; Levy and Sandler 2009), whereas audio content descriptors, including descriptors of timbre, are automatically extracted from audio (Lartillot and Toiviainen 2007; Peeters et al. 2011). Audio content descriptors for MPEG-7 include temporal (i.e., the root-mean-squared energy envelope, zero-crossing rate, temporal centroid, and autocorrelation coefficients), spectral (i.e., centroid, flatness, roll-off, and flux), cespstral (i.e., cepstral coefficients and MFCC), perceptual (i.e., sharpness), and specific descriptors (i.e., odd-to-even harmonic energy ratio, harmonic-noise ratio, and attack time).

The semantic tags in text-based descriptions commonly belong to a taxonomy that consists of a number of categories organized into a hierarchical tree used to provide semantic relationships between categories. For example, audio can be categorized into music, speech, or environmental sounds. Each of these categories can be further divided, such as the family of a musical instrument (i.e., brass, woodwinds, strings, and percussion), male or female speech, etc. As the taxonomy gets larger and more fully connected, the utility of the category relationships increases (Casey 2001b). Semantic tags are commonly used in text-based query applications, such as Internet search engines, where text from the query is matched against text from the tags (Casey 2001a). For example, the query "violin" would retrieve sounds tagged with "violin" and possibly also "musical instrument," "strings," etc. Query-by-example applications require audio content descriptors to retrieve sounds in a database that are similar to a target sound provided by the user. In this case, MPEG-7 audio content descriptors are used to compute the similarity with a distance metric such as dynamic time warping for hidden Markov models (Casey 2001b). *Hidden Markov models* are statistical models particularly suited to describe sequences where the probability of the current value depends on the previous value. In fact, Casey points out (also see Sect. 11.3.4) that sound phenomena are dynamic and the descriptors vary in time. In music and speech, this variation carries important information that plays a central role both in perception and in automated tasks. Thus, he proposes to use hidden Markov models in MPEG-7 sound recognition models. Hidden Markov models partition a sound class into a finite number of states, each of which is modeled by a continuous (typically Gaussian) probability distribution. Subsequently, individual sounds are described by their trajectories through this state space, also called *state path*. The state path is an important method of description in the context of MPEG-7 since it describes the evolution of a sound with respect to states that represent events such as onset, sustain, and release (Casey 2001b).

The categorical information in the MPEG-7 tags can be used for automatic classification in which the aim is to automatically assign a class from the taxonomy to audio to which the classifier has not had previous access. Automatic classification involves training statistical models to learn to recognize the class using a descriptor vector as input. Among the most widely used descriptors for automatic audio recognition and classification are representations derived from the power spectrum (Casey 2001a). The raw spectrum is rarely used as input in automatic classification due to the inherent high-dimensionality and redundancy. The typical number of bins of linearly spaced spectra lies between 128 and 1024, whereas probabilistic classifiers, such as hidden Markov models, commonly require low-dimensional representations, preferably fewer than 10 dimensions (Casey 2001b). In MPEG-7, the audio spectrum projection scheme (Casey 2001a; Kim et al. 2004) requires the application of dimensionality reduction techniques, such as principal component analysis or independent component analysis, prior to classification or query-by-example.

Casey concluded that MPEG-7 yielded very good recognizer performance across a broad range of sounds with applications in music genre classification. However, works that compared MFCC with MPEG-7 descriptors in multimedia indexing tasks, such as recognition, retrieval, and classification, found that MFCC outperformed MPEG-7.

Kim et al. (2004) compared the performance of MPEG-7 audio spectrum projection descriptors against MFCC in a video sound track classification task. They used three matrix decomposition algorithms to reduce the dimensionality of the MPEG-7 audio spectrum projection descriptors to 7, 13, and 23 dimensions and compared the resulting approaches with the same number of MFCC. They found that MFCC yielded better performance than MPEG-7 in most cases. They also pointed out that MPEG-7 descriptors are more computationally demanding to extract than MFCC.

Similarly, Deng et al. (2008) compared the performance of traditional descriptors (zero-crossing rate, root-mean-squared energy, spectral centroid, spectral spread, and spectral flux) with MFCC and MPEG-7 on automatic instrument classification tasks. They used several classification algorithms in musical instrument family classification, individual instrument classification, and classification of solo passages. Principal component analysis was used to reduce the dimensionality of MPEG-7 audio spectrum projection descriptors. They concluded that MFCC outperformed MPEG-7 and traditional descriptors when used individually.

Finally, Deng et al. (2008) tested descriptor combinations, such as MFCC with MPEG-7 and MFCC with traditional descriptors, and concluded that the addition of MPEG-7 to MFCC improved classification performance, whereas traditional descriptors plus MFCC yielded the poorest performance. They finally noted that the higher the dimensionality of the descriptors vector, the better the performance; so they tested the classification performance of descriptor combinations followed by dimensionality reduction with principal component analysis and found that the combinations exhibit strong redundancy.

MPEG-7 is a very ambitious international standard that encompasses audio, video, and multimedia description. MPEG-7 audio was developed to have a similar scale of impact on the future of music technology as the MIDI and MPEG-1 Audio Layer III (popularized as the MP3 format) standards have had in the past (Casey 2001b). However, more than 15 years after the introduction of the standard, the world of audio content descriptors still seems anything but standardized—researchers and practitioners continue to develop new approaches that are custom-made for the specific content-description problem at hand.

11.4.3 Computer-Aided Orchestration

Musical orchestration denotes the art of creating instrumental combinations, contrasts, and stratifications (see McAdams, Chap. 8). Initially, orchestration was restricted to the assignment of instruments to the score and, as such, was largely relegated to the background of the compositional process. Progressively, composers started regarding orchestration as an integral part of the compositional process whereby the musical ideas themselves are expressed. Compositional experimentation in orchestration originates from the desire to achieve musically intriguing timbres by means of instrumental combinations. However, orchestration manuals are notoriously empirical because of the difficulty in formalizing knowledge about the timbral result of instrument combinations.

Computer-aided orchestration tools (Carpentier et al. 2010a; Caetano et al. 2019) automate the search for instrument combinations that perceptually approximate a given reference timbre. The aim of computer-aided orchestration is to find a combination of notes from musical instruments that perceptually approximates a given reference sound when played together (Abreu et al. 2016; Caetano et al. 2019). Descriptors of timbre play a key role in the following steps of computer-aided orchestration: (1) timbre description of isolated sounds, (2) timbre description of combinations of musical instrument sounds, and (3) timbre similarity between instrument combinations and the reference sound.

The timbre of both the reference sound and of the isolated musical instrument sounds is represented with a descriptor vector comprising a subset of the traditional descriptors of timbre (Peeters et al. 2011). The extraction of the descriptors is computationally expensive, so the descriptors of the isolated musical instrument sounds are extracted prior to the search for instrument combinations and kept as metadata in a descriptor database. The descriptors for the reference sound are extracted for every new reference used.

Each instrument combination corresponds to a vector of descriptors that captures the timbral result of playing the instruments together. However, the total number of instrument combinations makes it impractical to extract descriptors for each possible combination (Carpentier et al. 2010a). Instead, the descriptor vector of an instrument combination is estimated from the descriptor vectors of the isolated sounds used in the combination (Carpentier et al. 2010b).

The timbral similarity between the reference sound and the instrument combination is estimated as the distance between the corresponding descriptor vectors. Smaller distances indicate a higher degree of timbral similarity (Carpentier et al. 2010a) with the reference, so the instrument combinations with the smallest distances are returned as proposed orchestrations for a given reference sound.

The resulting instrument combinations found to orchestrate a given reference sound will depend on which descriptors are included in the descriptor vector. For example, spectral shape descriptors focus on approximating the distribution of spectral energy of the reference sound. Carpentier et al. (2010a) proposed using the normalized harmonic energy, global noisiness, attack time, spectral flatness, roughness, frequency and amplitude of the energy modulation, and frequency and amplitude of the modulation of the fundamental frequency. Additionally, they added the following descriptors not related to timbre: fundamental frequency and total energy. Caetano et al. (2019) used the frequency and amplitude of the spectral peaks, spectral centroid, spectral spread, and also fundamental frequency, loudness, and root-mean-squared energy.

The type of reference sound to be orchestrated also plays a fundamental role in the instrument combinations found by computer-aided orchestration algorithms. For example, if the composer uses a clarinet sound as reference (and the musical instrument sound database contains clarinet sounds), the composer should naturally expect an isolated clarinet note to be the closest instrument combination found (unless the composer imposes constraints to the search such as returning instrument combinations without clarinet sounds or with at least three different instruments). Aesthetically interesting results can be achieved by choosing a reference sound that belongs to a different abstract category than musical instruments, such as environmental sounds or vocal utterances, because these references usually result in complex instrument combinations. To hear examples of orchestrations using different types of reference sounds, go to the sound files “car_horn.mp3”, “carnatic.mp3”, “choir.mp3”, and wind_harp.mp3”. Each sound file consists of the reference sound followed by four proposed orchestrations from Caetano et al. (2019).

Perceptually, two important phenomena contribute to attaining aesthetically interesting orchestrations: *timbre blends* and *sensory proximity*. Timbre blends occur when the timbre of the different instruments used in the combination fuse into a single percept (see McAdams, Chap. 8). The categorical distinction between the musical instruments must disappear so the sensory attributes of the combination emerge as a new timbre. Computer-aided orchestration algorithms find instrument combinations whose sensory attributes approximate those of the reference sound to evoke abstract auditory experiences. Audio descriptors of timbre play a key role in determining the timbre similarity between the instrument combinations and the reference sound (Siedenburg et al. 2016a). Traditionally, timbre similarity approaches compare time-averaged descriptor vectors from different musical instrument sounds, neglecting temporal variations (Esling and Agon 2013). While this is consistent with static timbre spaces, dynamic representations, such as the one shown in Fig. 11.7, require the use of time series of descriptors.

Computer-aided orchestration exemplifies the benefit of incorporating temporal information into timbre similarity. The static timbre similarity measure is appropriate when orchestrating reference sounds that are relatively stable (Carpentier et al. 2010a; Abreu et al. 2016). However, matching targets with dynamic variations, such as an elephant trumpeting, requires a time-series method that takes temporal variations of descriptors into consideration. Esling and Agon (2013) proposed a multi-objective time series-matching algorithm capable of coping with the temporal and multidimensional nature of timbre. The multi-objective time series-matching algorithm adopts a multi-dimensional measure of similarity that simultaneously optimizes the temporal evolution of multiple spectral properties and returns a set of efficient solutions rather than a single best solution.

11.4.4 Musical Instrument Sound Morphing

The aim of sound morphing is to synthesize sounds that gradually blur the categorical distinction between the sounds being morphed by blending their sensory attributes (Caetano and Rodet 2013). Therefore, sound morphing techniques allow synthesizing sounds with intermediate timbral qualities by interpolating the sounds being morphed. Fig. 11.8 shows a striking example of image morphing to illustrate sound morphing with a visual analogy. The morph is determined by a single parameter α that varies between 0 and 1. Only the source sound S is heard when $\alpha = 0$, whereas only the target sound T is heard when $\alpha = 1$. Intermediate values of α should correspond to perceptually intermediate sounds. However, simple morphing techniques seldom satisfy this perceptual requirement (Caetano and Rodet 2013). Sound morphing typically comprises the following steps: (1) modeling, (2) interpolation, and (3) resynthesis.

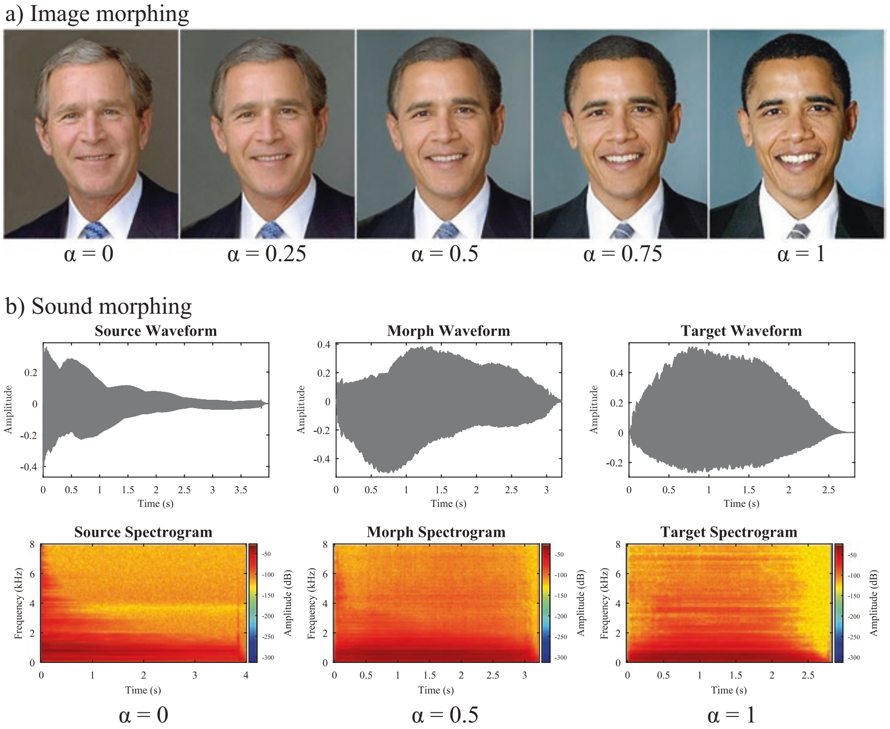

Fig. 11.8 Illustration of morphing for images and sounds: (a) face morphing; (b) musical instrument sound morphing. The source sound is the C#3 note played *forte* on a harpsichord and the target sound is the same note played *forte* on a tuba. The figure shows the *morphing factor* **α** below each corresponding panel. To hear the sounds, go to the sound file "harpsichord_tuba_morph.mp3". The image in (a) is currently publicly available at https://paulbakaus.com/wp-content/uploads/2009/10/bush-obama-morphing.jpg

The sounds being morphed (*S* and *T*) are modeled (e.g., with the sinusoidal model or the excitation-filter model) to obtain a parametric representation of *S* and *T*. For example, the parameters of the sinusoidal model are the frequencies and the amplitudes of the time-varying sinusoids that represent the partials of *S* and *T*. The parameters of the spectral envelope represent the filter component of the excitation-filter model. Cepstral coefficients and LPC are common representations of the filter in the excitation-filter model (Caetano and Rodet 2013).

The parameters of the morphed sound are obtained via linear interpolation between the parameters of *S* and *T*, for example, interpolation of the amplitudes and frequencies of the sinusoidal model or interpolation of the cepstral coefficients representing the spectral envelope of the excitation-filter model.

Finally, the morphed sound is resynthesized from the interpolated parameters. Perceptually, the model parameters play a crucial role in the final result, depending on the information captured by the model. For example, morphing with the sinusoidal model will result in intermediate amplitudes and frequencies (the model parameters), whereas morphing with the excitation-filter model will result in intermediate spectral envelopes.

The importance of the parametric representation is twofold: resynthesis and transformation. A parametric model should allow resynthesizing a sound that is perceptually very similar to the original sound from the model parameters alone. Sound transformations are achieved via manipulation of the model parameters followed by resynthesis, resulting in a sound that is perceptually different from the original sound. Striking transformations can be achieved by careful manipulation of model parameters depending on what information they represent. For example, the frequencies of the sinusoidal model can be manipulated to obtain a pitch transposition. Sound morphing is the result of parameter interpolation. However, most morphing techniques in the literature interpolate the parameters of the model used to represent the sounds regardless of the perceptual impact of doing so. Consequently, the morph is intermediate in the space of parameters rather than perceptually intermediate.

Caetano and Rodet (2013) used descriptors of timbre to guide musical instrument morphing toward more gradual transformations. They developed a sophisticated morphing technique based on a hybrid excitation-filter model where the filter is represented with spectral envelopes and the excitation has a sinusoidal component accounting for the partials and a residual component accounting for transients and noise missed by the sinusoids. Caetano and Rodet (2013) investigated the result of interpolating several representations of the spectral envelope: the spectral envelope curve, cepstral coefficients, LPC, reflection coefficients, and line spectral frequencies. Both reflection coefficients and line spectral frequencies arise from an interconnected tube model of the human vocal tract. Reflection coefficients represent the fraction of energy reflected at each section of the model, whereas line spectral frequencies represent the resonance conditions that describe the vocal tract being fully open or fully closed at the glottis (McLoughlin 2008).

Caetano and Rodet (2013) were interested in measuring the linearity of the morphing transformation with the different spectral envelope representations. They varied α linearly between 0 and 1 for each spectral envelope representation and recorded the corresponding variation of spectral shape descriptors (spectral centroid, spectral spread, spectral skewness, and spectral kurtosis). They found that linear interpolation of line spectral frequencies led to the most linear variation of spectral shape descriptors. Next, they performed a listening test to evaluate the perceptual linearity of the morphs with their hybrid excitation-filter model and the sinusoidal model. The listening test confirmed that the hybrid excitation-filter model resulted in morphs that were perceived as more perceptually linear than the sinusoidal model. Fig. 11.8 shows an example of musical instrument sound morphing from Caetano and Rodet (2013). To hear the sounds used in Fig. 11.8, go to the sound file "harpsichord_tuba_morph.mp3".

Perceptually, sound morphing can be viewed as an auditory illusion that is inherently intertwined with timbre because morphing manipulates both the sensory and the categorical perceptions of the sounds being morphed. For the sake of simplicity, the following examples will consider musical instruments and timbre spaces. In theory, sound morphing can break the categorical perception of musical instrument timbre. For example, when S and T are from different musical instruments, setting $\alpha = 0.5$ would produce a morph that theoretically resembles the sound of a hybrid musical instrument. Additionally, sound morphing can be used to create a *sonic continuum*. Timbre spaces are inherently sparse, with musical instrument sounds occupying specific points in an otherwise void space. Morphing musical instrument sounds can theoretically fill the gaps and create continuous timbre spaces by connecting musical instruments with intermediate sounds that no acoustical instrument can produce.

11.5 Summary

This chapter introduced the acoustic modeling of timbre via audio content descriptors. Sections were organized around the descriptor extraction process, covering important topics from general audio representations used to extract timbre descriptors to applications of these descriptors in psychology, sound synthesis, and music information retrieval. Audio content descriptors have played an important role in understanding the psychoacoustics of timbre, have become part of the industry standard MPEG-7 for audio content description, and play crucial roles for current developments of techniques such as computer-aided orchestration and musical instrument sound morphing. In these applications, audio descriptors help extract properties from the audio signal that are often of perceptual relevance and much more specific when compared to the general audio representations from which they are computed. At the same time, the audio descriptors described in this chapter are versatile enough to be valuable across a variety of different timbre-related audio processing tasks.

Audio descriptors could play a pivotal role in future research into timbre perception and sound processing in myriad ways. Section 11.4.1 outlined the ways in which the perception of timbral brightness has been modeled on acoustic grounds using audio descriptors. However, a model of timbre brightness perception that clearly delineates the acoustic ingredients of this important aspect of timbre perception has yet to be constructed and evaluated. Such a model would need to account for a variety of experimental phenomena (see McAdams, Chap. 2) across a large set of sounds. Section 11.4.3 and 11.4.4 summarized the role of audio descriptors in computer-aided orchestration and sound morphing. Here audio descriptors act as a perceptual proxy to allow synthesizing sounds with predefined perceptual characteristics. Adaptive processing (Verfaille et al. 2006) and content-based transformations (Amatriain et al. 2003) use audio descriptors to address the highly nonlinear connection between the audio and sound perception. However, the fundamental problem of synthesizing a waveform that matches a desired perceptual result remains a challenge.

Currently, the status of various approaches to audio content description is at a crossroads. The rise of machine learning architectures, such as deep neural networks, renders traditional audio descriptors obsolete in tasks such as musical instrument identification, environmental scene classification, or speaker recognition. Traditional audio descriptor-based classification architectures require two steps prior to learning per se: descriptor extraction followed by either descriptor selection or dimensionality reduction (see Sect. 11.3.4). One problem of these architectures is that they often fail to capture the highly nonlinear relationships commonly found in complex classification tasks. Deep neural networks are feed-forward artificial neural networks with several layers of hidden units between inputs and outputs (Hinton et al. 2012). The depth of the network provides sufficient flexibility to represent the nonlinearities critical to a given task such that deep neural networks jointly learn the descriptors and the classifier (Takahashi et al. 2018).

However, the main challenge of deep learning architectures lies in their application in timbre acoustics, perception, and cognition. Kell et al. (2018) made a significant contribution when they presented a deep neural network optimized for both speech and music recognition tasks. The deep neural network performed as well as humans, exhibited error patterns that resembled those of humans, and outperformed a linear spectrotemporal filter model of auditory cortex in the prediction of fMRI voxel responses. Moreover, the trained network replicated aspects of human cortical organization and provided evidence of hierarchical organization within the auditory cortex, with intermediate and deep layers best predicting primary and nonprimary auditory cortical responses, respectively. Nonetheless, prediction is not identical to understanding. Even though a good model should predict future data, a model needs to be transparent in order to allow for proper theory building. Future work in this direction will be able to draw insightful connections between the pattern of oscillations carried by sound waves and the timbre that listeners extract from these waves.

Compliance with Ethics Requirements Marcelo Caetano declares that he has no conflict of interest.

Charalampos Saitis declares that he has no conflict of interest.

Kai Siedenburg declares that he has no conflict of interest.

References

Abreu J, Caetano M, Penha R (2016) Computer-aided musical orchestration using an artificial immune system. In: Johnson C, Ciesielski V, Correia J, Machado P (eds) Evolutionary and biologically inspired music, sound, art and design, lecture notes in computer science, vol 9596. Springer, Heidelberg, pp 1–16

Almeida A, Schubert E, Smith J, Wolfe J (2017) Brightness scaling of periodic tones. Atten Percept Psychophys 79(7):1892–1896

Amatriain X, Bonada J, Loscos À et al (2003) Content-based transformations. J New Music Res 32(1):95–114

Aucouturier J-J, Defreville B, Pachet F (2007) The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music. J Acoust Soc Am. https://doi.org/10.1121/1.2750160

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

Bogert BP, Healy MJR, Tukey JW (1963) The quefrency analysis of time series for echoes: cepstrum, pseudo autocovariance, cross-cepstrum and saphe cracking. In: Rosenblatt M (ed) Time series analysis. Wiley, New York, pp 209–243

Brown JC (1991) Calculation of a constant Q spectral transform. J Acoust Soc Am 89(1):425–434

Brown JC (1999) Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. J Acoust Soc Am 105(3). https://doi.org/10.1121/1.426728

Brown JC, Houix O, McAdams S (2001) Feature dependence in the automatic identification of musical woodwind instruments. J Acoust Soc Am 109(3):1064–1072. https://doi.org/10.1121/1.1342075

Brown JC, Puckette MS (1992) An efficient algorithm for the calculation of a constant q transform. J Acoust Soc Am 92(5):2698–2701

Burred JJ, Röbel A (2010) A segmental spectro-temporal model of musical timbre. In: Zotter F (ed) Proceedings of the 13th international conference on digital audio effects (DAFx-10). IEM, Graz

Burred JJ, Röbel A, Sikora T (2010) Dynamic spectral envelope modeling for timbre analysis of musical instrument sounds. IEEE Trans Audio Speech Lang Proc 18(3):663–674

Caclin A, McAdams S, Smith BK, Winsberg S (2005) Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones. J Acoust Soc Am 118:471–482

Caetano MF, Burred JJ, Rodet X (2010) Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues. In: Zoter F (ed) Proceedings of the 13th international conference on digital audio effects (DAFx-10). IEM, Graz

Caetano M, Rodet X (2013) Musical instrument sound morphing guided by perceptually motivated features. IEEE Trans Audio Speech Lang Proc 21(8):1666–1675

Caetano M, Zacharakis A, Barbancho I, Tardón LJ (2019) Leveraging diversity in computer-aided musical orchestration with an artificial immune system for multi-modal optimization. Swarm Evol Comput. https://doi.org/10.1016/j.swevo.2018.12.010

Carpentier G, Assayag G, Saint-James E (2010a) Solving the musical orchestration problem using multiobjective constrained optimization with a genetic local search approach. J Heuristics 16(5):681–714. https://doi.org/10.1007/s10732-009-9113-7

Carpentier G, Tardieu D, Harvey J et al (2010b) Predicting timbre features of instrument sound combinations: application to automatic orchestration. J New Mus Res 39(1):47–61

Casey M (2001a) MPEG-7 sound-recognition tools. IEEE Trans Circ Sys Video Tech 11(6):737–747

Casey M (2001b) General sound classification and similarity in MPEG-7. Organized Sound 6(2):153–164

Casey MA, Veltkamp R, Goto M et al (2008) Content-based music information retrieval: current directions and future challenges. Proc IEEE 96(4):668–696

Childers DG, Skinner DP, Kemerait RC (1977) The cepstrum: a guide to processing. Proc IEEE 65(10):1428–1443

Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28(4):357–366

Deng JD, Simmermacher C, Cranefield S (2008) A study on feature analysis for musical instrument classification. IEEE Trans Syst Man Cybern B Cybern 38(2):429–438

De Poli G, Prandoni P (1997) Sonological models for timbre characterization. J New Mus Res 26(2):170–197

Dolson M (1986) The phase vocoder: a tutorial. Comp Music J 10(4):14–27. https://doi.org/10.2307/3680093

Esling P, Agon C (2013) Multiobjective time series matching for audio classification and retrieval. IEEE Trans Audio Speech Lang Proc 21(10):2057–2072

Fletcher NH (1999) The nonlinear physics of musical instruments. Rep Prog Phys 62(5):723–764

Giordano BL, McAdams S, Zatorre RJ et al (2012) Abstract encoding of auditory objects in cortical activity patterns. Cereb Cortex 23(9):2025–2037

Glasberg BR, Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. Hear Res 47:103–138

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

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

Hajda J (2007) The effect of dynamic acoustical features on musical timbre. In: Beauchamp JW (ed) Analysis, synthesis, and perception of musical sounds. Springer, New York, pp 250–271

Handel S (1995) Timbre perception and auditory object identification. In: Moore BCJ (ed) Hearing, Handbook of perception and cognition, 2nd edn. Academic Press, San Diego, pp 425–461

Harris FJ (1978) On the use of windows for harmonic analysis with the discrete Fourier transform. Proc IEEE 66(1):51–83

Hartmann WM (1996) Pitch, periodicity, and auditory organization. J Acoust Soc Am 100(6):3491–3502

Herrera-Boyer P, Peeters G, Dubnov S (2003) Automatic classification of musical instrument sounds. J New Music Res 32(1):3–21

Hinton G, Deng L, Yu D et al (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig Proc Mag 29(6):82–97

Holighaus N, Dörfler M, Velasco GA, Grill T (2013) A framework for invertible, real-time constant-Q transforms. IEEE Trans Audio Speech Lang Proc 21(4):775–785

Horner AB, Beauchamp JW, So RH (2011) Evaluation of Mel-band and MFCC-based error metrics for correspondence to discrimination of spectrally altered musical instrument sounds. J Audio Eng Soc 59(5):290–303

Huq A, Bello JP, Rowe R (2010) Automated music emotion recognition: a systematic evaluation. J New Mus Res 39(3):227–244

Irino T, Patterson RD (1997) A time-domain, level-dependent auditory filter: the gammachirp. J Acoust Soc Am 101:412–419

Jaffe DA (1987a) Spectrum analysis tutorial, part 1: the discrete Fourier transform. Comp Music J 11(2):9–24

Jaffe DA (1987b) Spectrum analysis tutorial, part 2: properties and applications of the discrete Fourier transform. Comp Music J 11(3):17–35

Kell AJE, Yamins DLK, Shook EN et al (2018) A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron 98(3):630–644. https://doi.org/10.1016/j.neuron.2018.03.044

Kim HG, Burred JJ, Sikora T (2004) How efficient is MPEG-7 for general sound recognition? Paper presented at the 25th international Audio Engineering Society conference: metadata for audio. London, 17–19 June 2004

Krimphoff J, Mcadams S, Winsberg S (1994) Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique (Characterization of the timbre of complex sounds II Acoustic analysis and psychophysical quantification) J de Physique (J Phys) IV(C5):625–628

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

Levy M, Sandler M (2009) Music information retrieval using social tags and audio. IEEE Trans Multimedia 11(3):383–395

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

McLoughlin IV (2008) Review: line spectral pairs. Sig Proc 88(3):448–467

Makhoul J (1975) Linear prediction: a tutorial review. Proc IEEE 63(4):561–580

Martínez JM, Koenen R, Pereira F (2002) MPEG-7: the generic multimedia content description standard, part 1. IEEE MultiMedia 9(2):78–87

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

Martínez J, Perez H, Escamilla E, Suzuki MM (2012). Speaker recognition using mel frequency cepstral coefficients (MFCC) and vector quantization (VQ) techniques. In: Sánchez PB (ed) Proceedings of the 22nd international conference on electrical communications and computers. IEEE, Piscataway, p 248–251

McAdams S, Douglas C, Vempala NN (2017) Perception and modeling of affective qualities of musical instrument sounds across pitch registers. Front Psychol. https://doi.org/10.3389/fpsyg.2017.00153

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

McAulay R, Quatieri T (1986) Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans Acoust Speech Sig Proc 34(4):744–754

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

Nack F, Lindsay AT (1999) Everything you wanted to know about MPEG-7: part 2. IEEE MultiMedia 6(4):64–73

Ogg M, Slevc LR, Idsardi WJ (2017) The time course of sound category identification: insights from acoustic features. J Acoust Soc Am 142(6):3459–3473

On CK, Pandiyan PM, Yaacob S, Saudi A (2006). Mel-frequency cepstral coefficient analysis in speech recognition. Paper presented at the 2006 international conference on computing & informatics (ICOCI 2006). Kuala Lumpur, 6–8 June 2006

Patterson RD, Robinson K et al (1992) Complex sounds and auditory images. In: Cazals Y, Demany L, Horner K (eds) Auditory physiology and perception. Pergamon Press, Oxford, pp 429–446

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

Pollard HF, Jansson EV (1982) A tristimulus method for the specification of musical timbre. Acta Acust united Ac 51(3):162–171

Portnoff M (1980) Time-frequency representation of digital signals and systems based on short-time Fourier analysis. IEEE Trans Acoust Speech Sig Proc 28(1):55–69

Regnier L, Peeters G (2009) Singing voice detection in music tracks using direct voice vibrato detection. In: Chen LG, Glass JR (eds) Proceedings of the 2009 IEEE international conference on acoustics, speech and signal processing, Taipei, April 2009. IEEE, Piscataway, p 1685–1688

Rigaud F, David B (2013) A parametric model and estimation techniques for the inharmonicity and tuning of the piano. J Acoust Soc Am 133(5):3107–3118. https://doi.org/10.1121/1.4799806

Saitis C, Giordano BL, Fritz C, Scavone GP (2012) Perceptual evaluation of violins: a quantitative analysis of preference judgements by experienced players. J Acoust Soc Am 132:4002–4012

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

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

Siedenburg K, Jones-Mollerup K, McAdams S (2016b) Acoustic and categorical dissimilarity of musical timbre: evidence from asymmetries between acoustic and chimeric sounds. Front Psychol 6(1977)

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

Slawson W (1985) Sound color. University of California Press, Berkeley

Stevens SS, Volkman J, Newman E (1937) A scale for the measurement of the psychological magnitude of pitch. J Acoust Soc Am 8(3):185–190

Takahashi N, Gygli M, Van Gool L (2018) AENet: learning deep audio features for video analysis. IEEE Trans Multimedia 20(3):513–524

Terasawa H, Slaney M, Berger J (2005) The thirteen colors of timbre. In: proceedings of the 2005 IEEE workshop on applications of signal processing to audio and acoustics, new Paltz, October 2005. IEEE, Piscataway, p 323–326

Verfaille V, Zolzer U, Arfib D (2006) Adaptive digital audio effects (a-DAFx): a new class of sound transformations. IEEE Trans Audio Speech Lang Proc 14(5):1817–1831

Zwicker E (1961) Subdivision of the audible frequency range into critical bands (Frequenzgruppen). J Acoust Soc Am 33:248–248

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