The Timbre Toolbox: Extracting audio descriptors from musical signals

Geoffroy Peeters[a)]
Institut de Recherche et Coordination Acoustique/Musique (STMS-IRCAM-CNRS), 1 place Igor-Stravinsky, F-75004 Paris, France

Bruno L. Giordano
Centre for Interdisciplinary Research on Music Media and Technology (CIRMMT), Schulich School of Music, McGill University, 555 Sherbrooke Street West, Montréal, Quebéc H3A 1E3, Canada

Patrick Susini and Nicolas Misdariis
Institut de Recherche et Coordination Acoustique/Musique (STMS-IRCAM-CNRS), 1 place Igor-Stravinsky, F-75004 Paris, France

Stephen McAdams
Centre for Interdisciplinary Research on Music Media and Technology (CIRMMT), Schulich School of Music, McGill University, 555 Sherbrooke Street West, Montréal, Quebéc H3A 1E3, Canada

(Received 24 November 2010; revised 9 March 2011; accepted 12 March 2011)

The analysis of musical signals to extract audio descriptors that can potentially characterize their timbre has been disparate and often too focused on a particular small set of sounds. The Timbre Toolbox provides a comprehensive set of descriptors that can be useful in perceptual research, as well as in music information retrieval and machine-learning approaches to content-based retrieval in large sound databases. Sound events are first analyzed in terms of various input representations (short-term Fourier transform, harmonic sinusoidal components, an auditory model based on the equivalent rectangular bandwidth concept, the energy envelope). A large number of audio descriptors are then derived from each of these representations to capture temporal, spectral, spectrotemporal, and energetic properties of the sound events. Some descriptors are global, providing a single value for the whole sound event, whereas others are time-varying. Robust descriptive statistics are used to characterize the time-varying descriptors. To examine the information redundancy across audio descriptors, correlational analysis followed by hierarchical clustering is performed. This analysis suggests ten classes of relatively independent audio descriptors, showing that the Timbre Toolbox is a multidimensional instrument for the measurement of the acoustical structure of complex sound signals. [DOI: 10.1121/1.3642604]

PACS number(s): 43.66.Jh, 43.75.Yy, 43.64.Bt, 43.60.Cg [DD] Pages: 2902–2916

I. INTRODUCTION

There is a growing interest within several domains of research and technology in establishing the acoustical basis of musical timbre perception. The term "timbre" encompasses a set of auditory attributes of sound events in addition to pitch, loudness, duration, and spatial position. Psychoacoustic research has modeled timbre as a multidimensional phenomenon and represents its perceptual structure in terms of "timbre spaces." It is important to be able to derive reliable acoustical parameters from the audio signal that can serve as potential physical correlates (or audio descriptors) of these dimensions. Composers and computer musicians need control over these acoustical parameters for sound synthesis and computer-aided orchestration. In the field of music information retrieval, perceptually relevant timbre parameters are needed as indices for content-based search of targeted timbres in very large sound databases, as well as for automatic categorization, recognition, and identification schemes for musical instrument and environmental sounds (McAdams, 1993). Having a systematic approach to sound analysis that is oriented towards human perception is thus a crucial step in applying musical acoustic research to these problem areas. This article describes a set of audio analysis tools that have been developed to achieve this goal, using a number of different input representations of the audio signal and numerous audio descriptors derived from those representations. It also conducts an analysis of the redundancy of information across the set of audio descriptors so that researchers can systematically select independent descriptors for their analyses. As such, the Timbre Toolbox, written in the MATLAB programming language, aims to provide a unique tool for the audio research and musical acoustics communities.

One of the most fruitful approaches to timbre perception has used multidimensional scaling analysis of dissimilarity ratings on pairs of musical instrument sounds differing primarily in their timbres (Plomp, 1970; Wedin and Goude, 1972; Wessel, 1973; Miller and Carterette, 1975; Grey,

a)Author to whom correspondence should be addressed. Electronic mail: geoffroy.peeters@ircam.fr

 0001-4966/2011/130(5)/2902/15/$30.00

1977; Wessel, 1979; Krumhansl, 1989; Iverson and Krumhansl, 1993; McAdams *et al.,* 1995; Kendall *et al.,* 1999; Lakatos, 2000; Marozeau *et al.,* 2003). In most of these studies, qualitative interpretations of the perceptual dimensions involved examining various acoustic representations of the signals and using them in a descriptive fashion to "explain" the perceptual results. Grey and Gordon (1978) were among the first to try to establish quantitative correlations between the position along a perceptual dimension and a value along an acoustic dimension derived from the sound signal, spectral centroid in their case. We will call such parameters "audio descriptors."[1] Subsequent work by Iverson and Krumhansl (1993), Krimphoff *et al.* (1994), McAdams *et al.* (1995), and Lakatos (2000) made a more systematic attempt at explaining all perceptual dimensions of a given timbre space by correlating acoustic parameters with perceptual dimensions. This approach led (1) to models of timbral distance based on audio descriptors (Misdariis *et al.,* 1998; Peeters *et al.,* 2000), some of which were included in MPEG-7 (ISO/IEC, 2002); (2) to the development of a large set of descriptors for use in music information retrieval and music content analysis (Fujinaga, 1998; Martin *et al.,* 1998; Fujinaga and MacMillan, 2000; Herrera *et al.,* 2000; Rioux *et al.,* 2002; Peeters, 2004; Tindale *et al.,* 2004); and (3) to confirmatory studies in which sounds were synthesized with specific acoustic properties to see if they could be recovered perceptually (Caclin *et al.,* 2005; Marozeau and de Cheveigné, 2007). Thus the development of audio descriptors has furthered research on musical timbre from several vantage points.

Quantitative studies of musical timbre have relied on different methods for extracting descriptors of the sound signals. As a result, the literature in this field lacks an exhaustive standard for the acoustical characterization of the signals. One of the main consequences of this fact is a decrease in the comparability of results from different studies. In human perception studies, for example, it is not possible to firmly conclude whether diverging results from psychoacoustic studies of musical timbre are due to the effect of variability in the sound stimuli or in the algorithm used to extract the audio descriptors. Further, in the machine-learning literature, it is not easy to establish whether differences across studies in classification performance are caused by a change in the sound-descriptor system or by differences in the mathematics of the classification algorithms. A second consequence of the variety of approaches to acoustical characterization is that no single study adopts a truly exhaustive system for characterizing acoustical signals: different studies are indeed likely to focus on aspects of the acoustical information that seem most relevant to their concerns. As a result, it is not possible to assess whether our knowledge of the human processing of complex sounds truly captures the entire gamut of perceptually relevant sound parameters. Similarly, music information retrieval studies might not exploit the full information potential of the sound signals, and hence may not attain the best possible performance allowed by the chosen classification strategy.

The Timbre Toolbox implements several different classes of audio descriptors related to the spectral, temporal, spectrotemporal, and intensive properties of the signals. The majority of the implemented audio descriptors have proven useful in various timbre-related tasks, such as explaining perceptual dimensions, performing acoustic content-based search in sound databases, and performing automatic musical instrument classification. In this article, we use the Timbre Toolbox to analyze a large database of musical sounds, the McGill University Master Samples library (Opolko and Wapnick, 2006). We also assess the informational redundancy of the Timbre Toolbox descriptors within the analyzed corpus of musical signals based on their intercorrelations. The goal of this analysis is to quantify the similarity of the various descriptors, to estimate approximately the number of groups of statistically independent descriptors, to assess the extent to which between-descriptor similarities are affected by a change in two important parameters of the analysis pipeline (input representation and the descriptive statistic used to summarize the time-varying descriptors over the duration of a sound event), and to provide recommendations that future studies can follow to select among the implemented descriptors.

II. STRUCTURE OF THE AUDIO DESCRIPTOR ANALYSIS SYSTEM

A. Global organization

A system for the extraction of audio descriptors is usually organized according to the properties of the descriptors. We can distinguish three main properties of an audio descriptor: (1) the temporal extent over which the descriptor is computed (a specific region in time, such as the sustain, or the whole duration of a sound file), (2) the signal representation used to compute it (e.g., the waveform, the energy envelope or the short-term Fourier transform), and (3) the descriptor concept described by it (e.g., the description of the spectral envelope or the energy envelope over time). We discuss these three properties below.

The *temporal extent* denotes the segment duration over which the descriptor is derived. A descriptor can either directly represent the whole sound event (e.g., the Log-Attack-Time descriptor, because there is only one attack in a sound sample) or represent a short-duration segment inside the event (e.g., the time-varying spectral centroid, which is derived from a spectral analysis of consecutive short-duration segments of a sound, usually of 60 ms duration). Descriptors of the first group are called "global descriptors," and those of the second group are called "time-varying descriptors." Time-varying descriptors are extracted within each time frame of the sound and therefore form a sequence of values. In order to summarize the sequence in terms of a single value, we use descriptive statistics, such as minimum or maximum values, the mean or median, and the standard deviation or interquartile range (i.e., the difference between the 75th and 25th percentiles of the sequence of values). As such, the structure of an audio descriptor system usually separates the extraction of global descriptors (which are directly considered as the final results) from the extraction of time-varying descriptors (which are subsequently processed to derive the descriptive statistics).

Most work on audio descriptors uses similar algorithms but with variations in the extraction process. Indeed

descriptors such as the spectral centroid can be extracted from various *input signal representations*. In our case, we consider the following input representations: Fourier spectrum (magnitude and power scales), harmonic sinusoidal components, and the output of a model of auditory processing—the Equivalent Rectangular Bandwidth (ERB) model. Such systems are thus usually organized as a set of mathematical operators (e.g., the formula for spectral centroid), which are applied to an input signal representation. To the contrary, some descriptor concepts can only be applied to specific signal representations. An example of this is the inharmonicity coefficient, which can only be derived from a harmonic signal representation.

Finally, one can attempt to distinguish descriptors according to the *concept* described. For example the autocorrelation coefficients, spectral centroid, spectral spread, spectral kurtosis, spectral skewness, spectral flatness, and spectral crest are all related to the shape of the spectrum, although they use different signal representations for their computation. We did not attempt to organize the descriptors according to these shared concepts, because this is subject to controversy: is the spectral flatness more related to an energy description than to a harmonicity description?

Below we first explain the input representations used and then explain the various operators applied to them to derive the audio descriptors. In Table I, we summarize the audio descriptors, their dimensionalities, the abbreviation we use to refer to them in Sec. IV, and the input representation used to compute them.

B. Input representations

The input of the audio descriptor analysis system is an audio signal. In the following, we denote it by $s(n)$ where $n \in \mathrm{N}^+$ is the sample number, or by $s(t_n)$ where $t_n = n/sr$ is the time expressed in seconds corresponding to n and to a sampling rate sr. The duration of the audio signal is denoted by L_n when expressed in samples and by L_t when expressed in seconds. For the extraction of the audio descriptors we considered the four following representations of the audio signal $s(t_n)$: (1) the temporal energy envelope, (2) the short-term Fourier transform, (3) the output of an auditory model, and (4) sinusoidal harmonic partials.

1. Temporal Energy Envelope

The temporal envelope $e(t_n)$ of the audio signal $s(t_n)$ is derived from the amplitude of the analytic signal $s_a(t_n)$ given by the Hilbert transform of $s(t_n)$. This amplitude signal is then low-pass filtered using a third-order Butterworth filter with a cutoff frequency of 5 Hz. $e(t_n)$ has the same sampling rate and duration as that of $s(t_n)$.

2. Short-term Fourier Transform (STFT amplitude and STFT power)

The STFT representation is obtained using a sliding-window analysis over the audio signal $s(t_n)$. We use a Hamming analysis window of 23.2 ms duration with a hop size of 5.8 ms. In the following we denote the center of one analysis window by m when expressed in samples and by t_m when expressed in seconds. The amplitude spectrum of the STFT is then used as one of the representations in order to derive the audio descriptors. Two types of scales are tested for the amplitude: a linear scale (called "magnitude" hereafter) and squared amplitude (called "power" hereafter). In the following, we denote the frequency and amplitude of the bin $k \in \mathrm{N}^+$ obtained at frame t_m by $f_k(t_m)$ and $a_k(t_m)$, respectively. In the case of the STFT, because the hop size is equal to 5.8 ms, the sampling rate is lower than that of the temporal envelope $e(t_n)$. It is 172.26 Hz independently of the audio signal sampling rate.

3. Auditory model (ERB gam and ERB fft)

One can model the way sounds are analyzed in the peripheral auditory system with a bank of bandpass filters whose bandwidths depend on the center frequency, a notion related to the concept of "critical band" (CB), based partly on the results of masking experiments. The Bark scale was proposed by Zwicker (1961) to provide an estimation of the CB. Another concept, the Equivalent Rectangular Bandwidth (ERB) has been proposed by Moore and Glasberg (1983) for modeling auditory filters based on more recent findings. The ERB of a given filter is equal to the bandwidth of a perfect rectangular filter with similar area and height. Moore and Glasberg proposed an equation describing the value of the ERB as a function of center frequency. Consequently, the frequency spectrum of a sound is assumed to be partitioned into B adjacent ERB filters used for calculating the audio descriptors based on a peripheral auditory system representation. In the implementation used in the Timbre Toolbox, the number of bands B depends on the sampling rate of the audio signal: $B = 77$ for $sr = 96$ kHz, 77 for 44.1 kHz, 69 for 22 kHz, and 56 for 11 kHz. One version uses a bank of gammatone filters (Patterson *et al.*, 1992) followed by temporal smoothing. Because of the differences in duration of the impulse response, the total temporal smoothing depends on frequency. The other version uses an FFT that gives an identical temporal response for all channels (which is useful for computing the time-varying spectral descriptors, for example). Both have approximately the same frequency resolution. As for the STFT, we used a hop size of 5.8 ms for the computation of the ERB using FFT.

4. Sinusoidal harmonic partials (Harmonic)

An audio signal can be represented as a sum of sinusoidal components (or partials) [cf. McAulay and Quatieri (1986) or Serra and Smith (1990)] with slowly varying frequency and amplitude:

$$s(t_n) \simeq \sum_{h=1}^{H} a_h(t_n) \cos(2\pi f_h(t_n) + \phi_{h,0}(t_n)), \quad (1)$$

where $a_h(t_n)$, $f_h(t_n)$, and $\phi_{h,0}(t_n)$ are the amplitude, frequency, and initial phase of partial h at time t_n. Given the assumption of slowly varying amplitude and frequency, $a_h(t_n)$ and $f_h(t_n)$ are lowpass signals that can therefore be estimated using

TABLE I. Audio descriptors, corresponding number of dimensions, unit, abbreviation used as the variable name in the MATLAB code and input signal representation. Units symbols: - = no unit (when the descriptor is "normalized"); a = amplitude of audio signal; F = Hz for the Harmonic, STFTmag and STFTpower representations, and ERB-rate units for the ERBfft and ERBgam representations; $I = a$ for the STFTmag representation and a^2 for the STFTpow, ERBfft and ERBgam representations.

	Audio descriptor	Units	Abbreviation	Input representation
Global descriptors	Attack	s	Att	Temporal Energy Envelope
	Decay	s	Dec	
	Release	s	Rel	
	Log-Attack Time	log(s)	LAT	
	Attack Slope	a/s	AttSlope	
	Decrease Slope	log(a)/s	DecSlope	
	Temporal Centroid	s	TempCent	
	Effective Duration	s	EffDur	
	Frequency of Energy Modulation	Hz	FreqMod	
	Amplitude of Energy Modulation	a	AmpMod	
Time-varying descriptors	Autocorrelation (12 coefficients)	-	AutoCorr	Audio Signal
	Zero Crossing Rate	s^{-1}	ZcrRate	
	RMS-Energy Envelope	a	RMSEnv	Temporal Energy Envelope
	Spectral Centroid	F	SpecCent	STFTmagnitude (STFTmag) STFTpower (STFTpow) ERBfft (ERBfft) ERBgammatone (ERBgam) Harmonic
	Spectral Spread	F	SpecSpread	
	Spectral Skewness	-	SpecSkew	
	Spectral Kurtosis	-	SpecKurt	
	Spectral Slope	F^{-1}	SpecSlope	
	Spectral Decrease	-	SpecDecr	
	Spectral Rolloff	F	SpecRollOff	
	Spectro-temporal variation	-	SpecVar	
	Frame Energy	I	FrameErg	
	Spectral Flatness	-	SpecFlat	STFTmag, STFTpow, ERBfft, ERBgam
	Spectral Crest	-	SpecCrest	
	Harmonic Energy	a^2	HarmErg	Harmonic
	Noise Energy	a^2	NoiseErg	
	Noisiness	-	Noisiness	
	Fundamental Frequency	Hz	F0	
	Inharmonicity	-	InHarm	
	Tristimulus (3 coefficients)	-	TriStim	
	Harmonic Spectral Deviation	a	HarmDev	
	Odd to even harmonic ratio	-	OddEveRatio	

frame analysis: $a_h(t_m)$ and $f_h(t_m)$. For this, we use a Blackman window of 100 ms duration and a hop size of 25 ms. It should be noted that this window duration is larger than that used for the computation of the STFT. The reason for this is to obtain a better spectral resolution (separation between adjacent spectral peaks), which is required in order to be able to describe harmonics individually and to compute the related harmonic descriptors. In line with Krimphoff *et al.* (1994) and Misdariis *et al.* (1998), the number of partials H is set to 20. This value represents a trade-off, because for a 50 Hz fundamental frequency it covers the range from 20 to 1000 Hz and for a 1000 Hz signal it covers the range from 1000 to 20 000 Hz. This parameter can easily be changed in the Timbre Toolbox.

In our system, the sinusoidal model is used for the estimation of harmonic descriptors such as the tristimulus (Pollard and Jansson, 1982) or the odd-to-even harmonic ratio (Caclin *et al.*, 2005). These descriptors require that an order and a number be assigned to the partials (e.g., we need to know which partials are the three first harmonics and which are odd- or even-numbered harmonics). We thus need to define a reference partial, as well as the relation between the partials h and the reference partial. Because of this constraint, we cannot use a blind sinusoidal model such as one that will only estimate partials using partial tracking.

We use a harmonic sinusoidal model extended to the slightly inharmonic case (such as for piano sounds), i.e., partials $f_h(t_m)$ are considered as multiples of a fundamental frequency $f_0(t_m)$ or as an inharmonic deformation of a harmonic series. For this, we define an inharmonicity coefficient $\alpha \geq 0$. The content of the spectrum is now explained by partials at frequencies $f_h(t_m) = f_0(t_m)h\sqrt{1+\alpha h^2}$. In order to estimate the model, we first estimate the fundamental frequency at each frame t_m. In the Timbre Toolbox implementation, we use the algorithm proposed by Camacho and Harris (2008). Given that $f_0(t_m)$ is an estimate, we allow a departure from the estimated value, denoted $f_h(t_m) = (f_0(t_m) + \delta(t_m))h\sqrt{1+\alpha h^2}$. For a given frame t_m, we then look for the best values of $\delta(t_m)$ and α (α is presumed to be constant over frames) such that the energy of the spectrum is best explained. We therefore search for values of $\delta(t_m)$ and α in order to maximize $e_{tm}(\delta, \alpha)$ defined as

$$e_{t_m}(\delta,\alpha) = \sum_h X_{t_m}\left((f_0(t_m)+\delta(t_m))h\sqrt{1+\alpha h^2}\right), \quad (2)$$

where $X_{tm}(f)$ is the amplitude of the DFT at frequency f and time t_m.

5. Comments on relationship between sampling rate, pitch, and representation

It should be noted that, in our system, all window durations and hop sizes are defined in seconds and then converted to samples according to the sampling rate of the input audio signal. This guarantees that the same spectral resolution will be obtained whatever the sampling rate of the signal. However, the content of the representation itself will differ according to the sampling rate. This is because the upper frequency of the STFT depends on the sampling rate (it is equal to $f_{\max} = sr/2$). The same is true for the number of harmonic partials that one can observe given a sampling rate or the number of ERB bands. According to the fundamental frequency of the audio signal, some representations may also coincide in the output. For example, if the signal is purely harmonic (without any noise), the STFT and sinusoidal harmonic partial representations will give similar audio descriptors. Also for very high fundamental frequencies, only a few partials may exist below the Nyquist frequency, and the ERB output may be limited to a few bands. It is also possible that too few harmonics exist to compute the audio descriptors based on the sinusoidal harmonic model. Therefore, when using the Timbre Toolbox, one should always keep in mind the meaning of each representation and descriptor when interpreting the descriptor values.

III. DEFINITION OF AUDIO DESCRIPTORS

In this section, we define the audio descriptors as operators applied to the four representations presented above. This formulation corresponds to the MATLAB code provided in the Timbre Toolbox (available for download at http://recherche.ircam.fr/pub/timbretoolbox or http://www.cirmmt.mcgill.ca/research/tools/timbretoolbox). In Table I, we provide the list of all audio descriptors, their respective dimensionalities, the units in which they are expressed, and the applicability of a given signal representation to compute them.

A. Temporal parameters

1. Computations on the audio signal $s(t_n)$

The autocorrelation coefficients and zero-crossing rate are time-varying descriptors computed directly from $s(t_n)$. The computation is performed using a sliding-window analysis with a window duration of 23.2 ms with a hop size of 2.9 ms. Its sampling rate is therefore 344.53 Hz independently of the audio signal sampling rate.

a. Autocorrelation coefficients. The autocorrelation coefficients (Brown, 1998) represent the spectral distribution of the signal $s(t_n)$ in the time domain (the autocorrelation of a signal is the inverse Fourier Transform of the spectral energy distribution of the signal). It has been proven to provide a good description for classification (Brown *et al.*, 2001). From the autocorrelation, we keep only the first 12 coefficients ($c \in \{1,\ldots, 12\}$), expressed as

$$\mathrm{xcorr}(c) = \frac{1}{\mathrm{xcorr}(0)} \sum_{n=0}^{L_n - c - 1} s(n)s(n+c), \tag{3}$$

where L_n is the window length expressed in samples and c is the time lag of the autocorrelation expressed in samples. It should be noted that, by its mathematical definition, the autocorrelation coefficients depend on the sampling rate, because the distance between two successive n is equal to $1/sr$. It is the only descriptor of the toolbox that depends on the sampling rate.

b. Zero-crossing rate. The zero-crossing rate is a measure of the number of times the value of the signal $s(t_n)$ crosses the zero axis. This value tends to be small for periodic sounds and large for noisy sounds. In order to compute this descriptor, the local DC offset of each frame of the signal is first subtracted. The zero-crossing rate value at each frame is then normalized by the window length L_t in seconds.

2. Energy envelope descriptors

The log-attack-time, attack-slope, decrease-slope, temporal-centroid, effective-duration, and energy-modulation are "global" descriptors computed using the energy envelope $e(t_n)$. It should be noted that the log-attack-time and attack-slope descriptors correspond closely to descriptors proposed by Gordon (1987), Krimphoff (1993), Krimphoff *et al.* (1994), and Wright (2008). In order to accurately estimate them, one needs a robust estimation of the location of the attack segment of a sound. Here we propose a new method to estimate it.

a. Attack estimation. In order to estimate the start (t_{st}) and end (t_{end}) times of the attack, many algorithms rely on fixed thresholds applied to the energy envelope $e(t_n)$ of the signal [for example defining t_{st} as the first value for which $e(t_n)$ goes above 10% of the maximum of $e(t_n)$ and t_{end} as the moment of the maximum of $e(t_n)$]. When applied to real sounds, this method was found not to be robust.[2] In order to address this problem, we use the "weakest-effort method" proposed by Peeters (2004), in which the thresholds are not fixed but are estimated according to the behavior of the signal during the attack. We first define a set of thresholds $\theta_i = \{0.1, 0.2, 0.3,\ldots, 1\}$ as a proportion of the maximum of the energy envelope. For each threshold θ_i, we estimate the time t_i at which the energy envelope $e(t_n)$ reaches this threshold for the first time: t_i such that $e(t_i) = \theta_i \max(e(t_n))$. We then define "effort" as the time interval between two successive t_i, so named because it represents the effort taken by the energy to go from one threshold to the next: $\omega_{i,i+1} = t_{i+1} - t_i$. This is illustrated in Fig. 1. The average value of the "efforts" $\overline{\omega}$ is then computed. The best threshold to be used for the estimation of the start of the attack θ_{st} is then defined as the first θ_i for which the effort $\omega_{i,i+1}$ goes below the value $\alpha\overline{\omega}$ with $\alpha > 1$. In other words, we are looking for the first threshold for which the corresponding effort is "weak": it is $\omega_{2,3}$ in Fig. 1. In a similar way, the best threshold to be used for the estimation of the end of the attack θ_{end} is defined as the last θ_i for which the effort $\omega_{i,i+1}$ goes below the value $\alpha\overline{\omega}$. It is $\omega_{7,8}$ in Fig. 1. After experimenting on 1500 sounds from the Ircam Studio On Line instrument database, we have

set $\alpha = 3$. Finally, the exact start time (t_{st}) and end time (t_{end}) of the attack are estimated by taking the minimum and maximum values of $e(t_n)$ in the intervals $\omega_{i,i+1}$ corresponding to θ_{st} and θ_{end} ($\omega_{2,3}$ and $\omega_{7,8}$ in Fig. 1).

b. Log-attack-time. The log-attack-time is simply defined as

$$\mathrm{LAT} = \log_{10}(t_{end} - t_{st}). \tag{4}$$

c. Attack slope. The attack slope is defined as the average temporal slope of the energy during the attack segment. We compute the local slopes of the energy corresponding to each effort w_i. We then compute a weighted average of the slopes. The weights are chosen in order to emphasize slope values in the middle of the attack (the weights are the values of a Gaussian function centered around threshold = 50% and with a standard-deviation of 0.5).

d. Decrease slope. The temporal decrease is a measure of the rate of decrease of the signal energy. It distinguishes non-sustained (e.g., percussive, pizzicato) sounds from sustained sounds. Its calculation is based on a decreasing exponential model of the energy envelope starting from its maximum (t_{max}):

$$\hat{e}(t_n) = Ae^{-\alpha(t_n - t_{max})} t_n \max, \tag{5}$$

where α is estimated by linear regression on the logarithm of the energy envelope.

e. Temporal centroid. The temporal centroid is the center of gravity of the energy envelope. It distinguishes percussive from sustained sounds. It has been proven to be a perceptually important descriptor (Peeters *et al.*, 2000):

$$tc = \frac{\sum_{n=n_1}^{n=n_2} t_n \cdot e(t_n)}{\sum_{n} e(t_n)}, \tag{6}$$

where n_1 and n_2 are the first and last values of n, respectively, such that $e(t_n)$ is above 15% of its maximum value. This is used in order to avoid including silent segments in the computation of tc.

f. Effective duration. The effective duration is a measure intended to reflect the perceived duration of the signal. It distinguishes percussive sounds from sustained sounds but depends on the event duration. It is approximated by the time the energy envelope $e(t_n)$ is above a given threshold. After many empirical tests, we have set this threshold to 40%.

g. Energy modulation (tremolo). On the sustained part of the sound (the part used for the computation of the decrease slope), denoted by S, we represent the modulation of the energy over time using a sinusoidal component. We estimate the amplitude and frequency (in Hz) of the modulation. This representation corresponds roughly to a tremolo

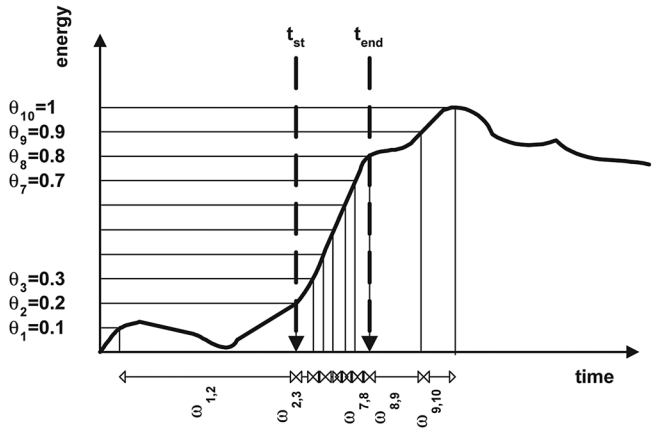

FIG. 1. Estimation of the attack segment using Peeters' (2004) weakest-effort method.

model. For this, we first subtract from the time trajectory of the energy $e(t_n \in S)$, the model $\hat{e}(t_n \in S)$ used for the computation of the decrease slope. The resulting residual signal is then analyzed using a DFT. The maximum peak of the DFT in the range 1 to 10 Hz is then estimated and is used as an estimate of the modulation amplitude and frequency. If no peak is detected, the modulation amplitude is set to 0.

B. Spectral parameters

All spectral parameters are time-varying descriptors computed using either the magnitude STFT, the power STFT, the harmonic sinusoidal partials or the ERB model output. In the following, $a_k(t_m)$ represents the value at bin k of the magnitude STFT, the power STFT, the $k = h$ sinusoidal harmonic partial or the k^{th} ERB filter. We denote the frequency (in Hz) corresponding to k by f_k. We define the normalized form of a_k by $p_k(t_m) = [a_k(t_m)]/\sum_{k=1}^{K} a_k(t_m)$. Therefore, $p_k(t_m)$ represents the normalized value of the magnitude STFT, the power STFT, sinusoidal harmonic partial or ERB filter at bin k and time t_m. p_k may be considered as the probability of observing k.

1. Frame energy

The frame energy is computed as the sum of the squared amplitudes (a_k^2) (being STFT or harmonic partials coefficients) at time t_m: $E_T(t_m) = \sum_k a_k^2(t_m)$. It should be noted that the window used to perform the frame analysis is normalized in amplitude such that its length or shape do not influence the value obtained.

2. Statistical moments of the spectrum

The following set of audio descriptors are the first four statistical moments of the spectrum.

Spectral centroid represents the spectral center of gravity. It is defined as

$$\mu_1(t_m) = \sum_{k=1}^{K} f_k \cdot p_k(t_m). \tag{7}$$

Spectral spread or spectral standard-deviation represents the spread of the spectrum around its mean value. It is defined as

$$\mu_2(t_m) = \left(\sum_{k=1}^{K} (f_k - \mu_1(t_m))^2 \cdot p_k(t_m) \right)^{1/2}. \quad (8)$$

Spectral skewness gives a measure of the asymmetry of the spectrum around its mean value. $\mu_3 = 0$ indicates a symmetric distribution, $\mu_3 < 0$ more energy at frequencies lower than the mean value, and $\mu_3 > 0$ more energy at higher frequencies:

$$\mu_3(t_m) = \left(\sum_{k=1}^{K} (f_k - \mu_1(t_m))^3 \cdot p_k(t_m) \right) \Big/ \mu_2^3. \quad (9)$$

Spectral kurtosis gives a measure of the flatness of the spectrum around its mean value. $\mu_4 = 3$ indicates a normal (Gaussian) distribution, $\mu_4 < 3$ a flatter distribution, and $\mu_4 > 3$ a peakier distribution

$$\mu_4(t_m) = \left(\sum_{k=1}^{K} (f_k - \mu_1(t_m))^4 \cdot p_k(t_m) \right) \Big/ \mu_2^4. \quad (10)$$

3. Description of the slope of the spectrum

The next set of descriptors is related to the slope of the spectrum.

Spectral slope is computed using a linear regression over the spectral amplitude values. It should be noted that the spectral slope is linearly dependent on the spectral centroid:

$$\text{slope}(t_m) = \frac{1}{\sum_{k=1}^{K} a_k(t_m)} \times \frac{K \sum_{k=1}^{K} f_k a_k(t_m) - \sum_{k=1}^{K} f_k \cdot \sum_{k=1}^{K} a_k(t_m)}{K \sum_{k=1}^{K} f_k^2 - \left(\sum_{k=1}^{K} f_k \right)^2}. \quad (11)$$

Spectral decrease was proposed by Krimphoff (1993) in relation to perceptual studies. It averages the set of slopes between frequency f_k and f_1. It therefore emphasizes the slopes of the lowest frequencies:

$$\text{decrease}(t_m) = \frac{1}{\sum_{k=2}^{K} a_k(t_m)} \sum_{k=2}^{K} \frac{a_k(t_m) - a_1(t_m)}{k-1}. \quad (12)$$

Spectral roll-off was proposed by Scheirer and Slaney (1997). It is defined as the frequency $f_c(t_m)$ below which 95% of the signal energy is contained:

$$\sum_{f=0}^{f_c(t_m)} a_f^2(t_m) = 0.95 \sum_{f=0}^{sr/2} a_f^2(t_m), \quad (13)$$

where $sr/2$ is the Nyquist frequency. In the case of harmonic sounds, it can be shown experimentally that spectral roll-off is related to the harmonic/noise cutoff frequency.

4. Description of the tonal/noise content of the spectrum

Spectral-flatness measure (SFM) and spectral-crest measures (SCM) have been proposed in the context of speech description (Johnston, 1988) and in the context of the MPEG-7 Audio standard (ISO/IEC, 2002). Under the assumption that a white noise produces a flat spectrum and that a sinusoidal component produces a peak in the spectrum, the measure of the flatness of the spectrum roughly discriminates noise from harmonic content.

The *spectral flatness measure* is obtained by comparing the geometrical mean and the arithmetical mean of the spectrum. The original formulation first split the spectrum into various frequency bands (Johnston, 1988). However, in the context of timbre characterization, we use a single frequency band covering the whole frequency range. For tonal signals, SFM is close to 0 (peaky spectrum), whereas for noisy signals it is close to 1 (flat spectrum):

$$\text{SFM}(t_m) = \frac{\left(\prod_{k=1}^{K} a_k(t_m) \right)^{1/K}}{\frac{1}{K} \sum_{k=1}^{K} a_k(t_m)}. \quad (14)$$

In the same spirit, the *spectral crest measure* is obtained by comparing the maximum value and arithmetical mean of the spectrum:

$$\text{SCM}(t_m) = \frac{\max_k^a k(t_m)}{\frac{1}{K} \sum_{k=1}^{K} a_k(t_m)}. \quad (15)$$

C. Parameters specific to the harmonic analysis

The following set of parameters are also time-varying descriptors but can only be computed using a sinusoidal harmonic partial representation. We denote by $a_h(t_m)$ and $f_h(t_m)$ the amplitude and frequency of partial h at time t_m. We estimate H partials ranked by increasing frequency.

1. Parameters related to the energy content

Harmonic energy is the energy of the signal explained by the harmonic partials. It is obtained by summing the energy of the partials detected at a specific time t_m:

$$E_H(t_m) = \sum_{h=1}^{H} a_h^2(t_m). \quad (16)$$

Noise energy is the energy of the signal not explained by harmonic partials. We approximate it by subtracting the harmonic energy from the total energy:

$$E_N(t_m) = E_T(t_m) - E_H(t_m). \quad (17)$$

Noisiness is the ratio of the noise energy to the total energy:

$$\text{noisiness}(t_m) = \frac{E_N(t_m)}{E_T(t_m)}. \quad (18)$$

High noisiness values indicate a signal that is mainly nonharmonic.

The *tristimulus* values were introduced by Pollard and Jansson (1982) as a timbral equivalent to color attributes in vision. The tristimulus comprises three different energy ratios allowing a fine description of the first harmonics of the spectrum:

$$T1(t_m) = \frac{a_1(t_m)}{\sum_{h=1}^{H} a_h(t_m)},$$

$$T2(t_m) = \frac{a_2(t_m) + a_3(t_m) + a_4(t_m)}{\sum_{h=1}^{H} a_h(t_m)},$$

$$T3(t_m) = \frac{\sum_{h=5}^{H} a_h(t_m)}{\sum_{h=1}^{H} a_h(t_m)}, \quad (19)$$

where H is the total number of partials considered (by default $H = 20$ in the Timbre Toolbox).

2. Parameters related to the frequency content

The *fundamental frequency*, denoted by $f_0(t_m)$, can be estimated using the algorithm of Maher and Beauchamp (1994) or de Cheveigné and Kawahara (2002). In the Timbre Toolbox, we use the algorithm of Camacho and Harris (2008).

Inharmonicity measures the departure of the frequencies of the partials f_h from purely harmonic frequencies hf_0. It is estimated as the weighted sum of deviation of each individual partial from harmonicity:

$$\text{inharmo}(t_m) = \frac{2}{f_0(t_m)} \frac{\sum_{h=1}^{H} (f_h(t_m) - hf_0(t_m)) a_h^2(t_m)}{\sum_{h=1}^{H} a_h^2(t_m)}. \quad (20)$$

Harmonic spectral deviation measures the deviation of the amplitudes of the partials from a global (smoothed) spectral envelope (Krimphoff *et al.*, 1994):

$$\text{HDEV}(t_m) = \frac{1}{H} \sum_{h=1}^{H} (a_h(t_m) - SE(f_h, t_m))), \quad (21)$$

where $SE(f_h, t_m)$ denotes the value of the spectral envelope at frequency f_h and time t_m. The spectral envelope at frequency f_h can be roughly estimated by averaging the values of three adjacent partials:

$$\text{SE}(f_h, t_m) = \frac{1}{3}(a_{h-1}(t_m) + a_h(t_m) + a_{h+1}(t_m)) \; for \; 1 < h < H. \quad (22)$$

The *odd-to-even harmonic energy ratio* distinguishes sounds with predominant energy at odd harmonics (such as clarinet sounds) from other sounds with smoother spectral envelopes (such as the trumpet):

$$\text{OER}(t_m) = \frac{\sum_{h=1}^{H/2} a_{2h-1}^2(t_m)}{\sum_{h=1}^{H/2} a_{2h}^2(t_m)}. \quad (23)$$

D. Spectro-temporal parameters

Spectral variation (also called spectral flux) is a time-varying descriptor computed using either the magnitude STFT, the power STFT, the harmonic sinusoidal partials or the ERB model output. It represents the amount of variation of the spectrum over time, defined as 1 minus the normalized correlation between the successive a_k (or a_h in the case of the harmonic sinusoidal model) (Krimphoff *et al.*, 1994):

$$\text{variation}(t_m, t_m - 1) = 1 - \frac{\sum_{k=1}^{K} a_k(t_{m-1}) a_k(t_m)}{\sqrt{\sum_{k=1}^{K} a_k(t_{m-1})^2} \sqrt{\sum_{k=1}^{K} a_k(t_m)^2}}. \quad (24)$$

E. Descriptive statistics of time-varying descriptors

We denote by $D(t_m)$ the value of a specific time-varying audio descriptor D at frame t_m. In order to summarize as a single value properties of the sequence of values $D(t_m)$, we apply a set of descriptive statistics. Among the statistics commonly used (minimum, maximum, mean, median, standard deviation, and interquartile range), we consider only the median as a measure of central tendency and the interquartile range as a measure of variability. Indeed, when analyzing a real audio signal, part of the values of $D(t_m)$ can potentially correspond to a silent segment in the signal. The corresponding values, apart from being meaningless for describing the timbre of the sound, will constitute outliers and dramatically influence the computations of mean, standard-deviation, min and max values. One could apply a threshold based on the loudness level to avoid that, but this would necessitate the definition of a threshold level, which can be problematic in terms of generalization across sounds and sound sets. For this reason, we employ the more robust measures of median and interquartile range. It should be noted that in the case in which $D(t_m)$ follows a normal distribution, the median is equal to the mean and the interquartile range is 1.349 times the standard deviation.

IV. ANALYSIS OF THE INDEPENDENCE OF AUDIO DESCRIPTORS

One of the main aims of this article is to assess the redundancy of the information quantified by each of the descriptors in relation to the other descriptors in the Timbre Toolbox. To this purpose, we focused on the correlations among descriptors, where pairs of descriptors characterized by a large absolute correlation also share a large amount of information concerning the sound signal.

In a first analysis, we compared the extent to which a choice of input representation and a choice of descriptive statistic for time-varying descriptors affects the structure of the correlations among the audio descriptors. Part of the interest in this analysis was practical. Timbre researchers may face a choice among input representations, some of which can be putatively better adapted to the purpose of the study (e.g., better modeling of peripheral auditory processes) yet more expensive from a computational point of view. Within this context, a choice among input representations might benefit from knowledge of the effect of a change in input representation on the structure of the correlations among descriptors. Indeed, if various input representations yield highly similar structures for the between-descriptor correlations, they will likely perform similarly in a variety of tasks such as explaining the ratings of participants in a timbre perception study, for example. A similar logic concerns choices made among the descriptive statistics used to summarize time-varying descriptors.

In a second analysis, we created two models that describe the correlational distances among the audio descriptors implemented in the Timbre Toolbox: a hierarchical clustering solution and a multidimensional scaling (MDS) representation. These representations can provide useful constraints to researchers faced with the problem of selecting among the descriptors in the database. Indeed, swamping an empirical investigation with an exceedingly high number of descriptors can prove counterproductive if they are strongly correlated. Although further analyses focused on the hierarchical clustering model, the MDS representation is presented here in order to provide a complementary visualization of the structure of the between-descriptor correlations.

The goal of a final analysis was to estimate approximately the number of groups of independent descriptors implemented in the Timbre Toolbox as derived from the cluster analysis. On the one hand, this additional analysis complements the distance models of the between-descriptor correlations and can thus provide more specific guidelines in the process of descriptor selection. On the other hand, the number of groups of statistically independent descriptors provides an estimate of the informational richness of the Timbre Toolbox itself with respect to one large musical sound database.

Data-analytic procedures were designed to extract trends that are valid across many dimensions of musical variation (e.g., pitch, duration, dynamics). To this purpose, (1) we adopted statistical tools that are robust, i.e., that measure trends in the majority of the datapoints without being influenced by outliers; (2) we focused on the most robust descriptive statistics for the time-varying descriptors among those available in the Timbre Toolbox, the median and interquartile range (IQR); and (3) we analyzed a large database of highly heterogenous signals comprising musical sounds similar to those frequently investigated in past studies, as well as additional musical signals of potential interest to future studies on musical timbre.

A. Sound samples

We analyzed a large portion of the musical signals from the McGill University Master Samples (MUMS) database (Opolko and Wapnick, 2006, sampling rate = 44.1 kHz; bit-depth = 16 bit). The selection criteria focused on sounds most commonly investigated in perceptual, music-retrieval and machine-classification studies of musical timbre. In particular, we considered sound samples that (1) contained a single tone of a single musical instrument, excluding chords, ensemble sounds where different musical instruments played tones at different pitches, glissandi, and percussion patterns and (2) were not the product of audio effects rarely adopted in music performance (e.g., pitch-shifted tom-tom sound). Based on these criteria, we selected 6037 sound samples, more than 90% of the samples in the database. The sample set comprised multiple instruments from each of the families of musical instruments: aerophones (wind instruments), chordophones (string instruments), membranophones, idiophones, and electrophones. They included pitches from A0 to G#8 (27.5-6645 Hz; median pitch = F4, 349.3 Hz, see Fig. 2) and several nonpitched percussive samples. The samples also covered a large range of durations (from hundreds of milliseconds up to 8 s ca.), as estimated from their effective duration, and a large dynamic range ($\approx$ 55 dB), as estimated from the median of the time-varying STFTmag Frame Energy (see Fig. 2).

The sounds are presented in a continuous fashion in the MUMS database, and the exact onsets and offsets are not indicated. In order to extract individual sounds for analysis, we were therefore obliged to estimate the onsets and offsets based on an analysis of the temporal envelope $e(t)$. The envelope for each sound was derived by forward-reverse low-pass filtering of the Hilbert transform of the waveform (third-order Butterworth, cut-off frequency = 20 Hz). Note that the extraction of $e(t)$ as described in Sec. II B 1 uses one simple filtering step and adopts a lower cut-off frequency (5 Hz). Here, we used a higher cut-off frequency and forward-reverse filtering to compensate for the delays in the IIR filter and to achieve a more accurate estimation of onset and offset for rapidly varying signals (e.g., percussion sounds). Starting from the peak of the temporal envelope and moving backwards in time, onset was defined as the temporal position of the last $e(t)$ sample whose level was within 35 dB of the peak level. Starting from the peak and moving forward, offset was defined as the time of the first $e(t)$ sample whose level was lower than 35 dB relative to the peak level.

B. Intercorrelations among audio descriptors

We carried out three different analyses: (1) comparison of the effects of the different input representations and descriptive statistics of the time-varying descriptors on the correlation distances among descriptors; (2) development of distance models for the matrix of between-descriptor correlations; and (3) assessment of the number of groups of independent audio descriptors.

In the following, *ad* refers to the ensemble of values for a particular audio descriptor within the analyzed database of musical signals, and r and t refer to the input representation and the descriptive statistic used to summarize quantitatively a time-varying audio descriptor, respectively. For all analyses, the pairwise correlational distance $d(ad_{rt}, ad'_{r't'})$ between

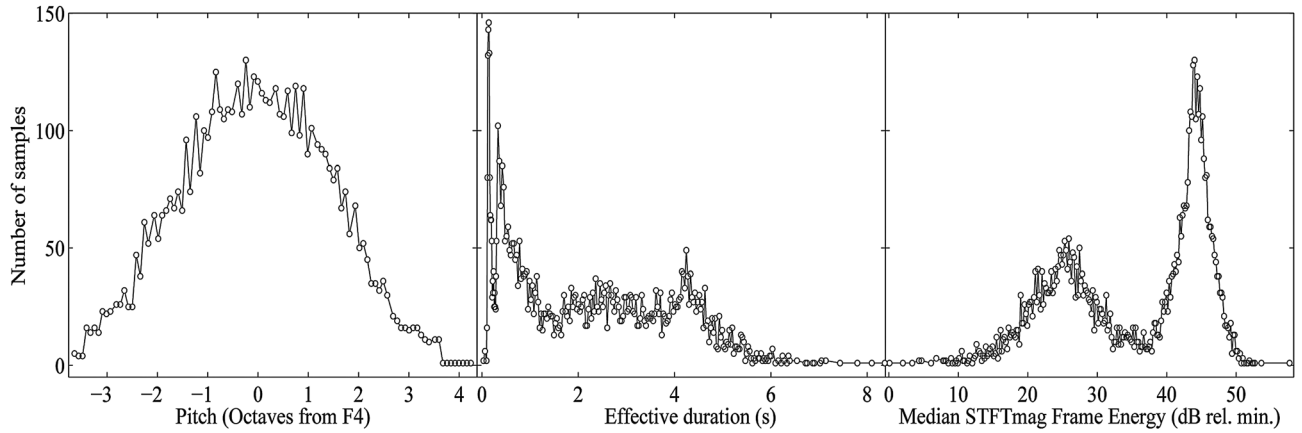

FIG. 2. Distribution of three non-timbre acoustical properties in the MUMS database (F0, duration and energy), as estimated with the Timbre Toolbox.

descriptor ad_{rt} and $ad'_{r't'}$ was defined as $D = 1 - |\rho_S(ad_{rt}, ad'_{r't'})|$ where ρ_S is a robust Spearman rank correlation. We chose rank correlation over the linear Pearson correlation because it is able to capture associations independent of monotone transforms (e.g., the rank correlation between the fundamental frequency and the log of the fundamental frequency equals one when focusing on ranks, whereas the Pearson correlation between the same variables is lower than one). With all analyses, we collapsed correlation distances across the coefficients of the multicoefficient descriptors autocorrelation and tristimulus. To this purpose, the collapsed absolute correlation distance $D(ad_t, ad'_{r't'}) = \text{Median}(D(ad_{it}, ad'_{r't'}))$ where $ad_{it} =$ autocorrelation or tristimulus, and i indexes the coefficients of ad_{it}. For analysis (1), this simplification avoided an overemphasis of the effects of the descriptive statistics for the autocorrelation and tristimulus descriptors. For analysis (2), this simplification reduced the complexity of the models of the between-descriptor distances, while still capturing the major trends in the correlations between the autocorrelation and tristimulus descriptors on the one hand, and the rest of the descriptors in the Timbre Toolbox on the other.

As can be noted in Table I, several of the descriptors in the Timbre Toolbox can be computed based on different representations r and/or descriptive statistics t (e.g., for Spectral Flatness $r = \{$STFTmag, STFTpow, ERBfft, ERBgam, Harmonic$\}$ and $t = \{$Med, IQR$\}$; for Noisiness $r = \{$Harmonic$\}$ and $t = \{$Med, IQR$\}$). The goal of an initial analysis was thus to assess the extent to which choosing among $r = \{$STFTmag, STFTpow, ERBfft, ERBgam, Harmonic$\}$ and $t = \{$Med, IQR$\}$ affected the correlation distances among all the descriptors in the Timbre Toolbox, including those that could be extracted based on only one input representation (e.g., harmonic energy for which $r = \{$Harmonic$\}$ and $t = \{$Med, IQR$\}$, and the global descriptor attack, for which $r =$ temporal energy envelope).

To this purpose, we computed ten different matrices $D(r, t)$ of the correlation distances among descriptors by combining factorially the five possible choices of r with the two possible choices of t. In the case of the D(STFTmag, IQR) matrix we considered the following: (1) the IQR descriptors extracted from STFTmag but not, for example, from STFTpow; (2) the IQR descriptors that can be extracted from only one input representation, such as zero-crossing rate or harmonic energy; and (3) the global descriptors extracted from the temporal energy envelope.

The next step of this analysis was to quantify the changes in the between-descriptor correlation distances created by a change in r and t. To this purpose, we computed a measure of the pairwise distance Δ between the ten D matrices defined as $\Delta = 1 - \rho_P(D(r, t), D(r', t'))$, where ρ_P is the robust Pearson correlation between the lower triangular portion of the correlation matrices $D(r, t)$ and $D(r', t')$. Figure 3 shows an agglomerative hierarchical clustering model of the distance matrix Δ (median linkage; cophenetic correlation between Δ and hierarchical-clustering distance $= 0.9$). A change in input representation r leads to changes in the between-descriptor distances $D(r, t)$ that are smaller than those resulting from a change in descriptive statistic t. For example, D(STFTmag, IQR) is more strongly correlated with D(ERBgam, IQR) than with D(STFTmag, Med). Very similar results, not reported here for the sake of brevity, were obtained when ρ_P was computed by ignoring elements common to all the distance matrices D (e.g., all of them

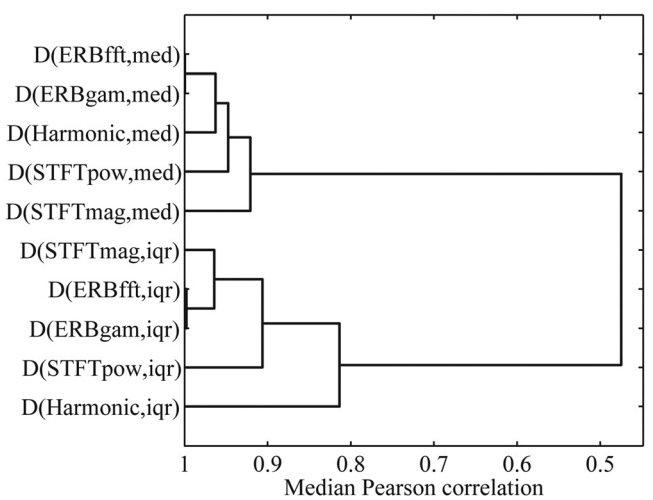

FIG. 3. Effect of the choice of input representation and descriptive statistic for time-varying descriptors on the correlation distance D between descriptors in the Timbre Toolbox. Med = median; iqr = interquartile range.

contained the same correlations between global descriptors derived from the temporal energy envelope), and when defining Δ as the Euclidean distance between D matrices.

In a second analysis, we modeled the between-descriptor distances D using two distance models: hierarchical clustering and metric MDS. In order to reduce the complexity of the distance models, we collapsed correlation distances across variants of the same descriptor based on the same descriptive statistic but on different input representations, i.e., the collapsed correlation distance $D'(ad_t, ad'_{t'}) = \text{Median}(D(ad_{rit}, ad'_{r'jt'}))$ where i and j index the input representations allowed for descriptor ad and ad', respectively, and $i = j$ if both ad and ad' can be computed using the same set of input representations. The choice of collapsing correlation distances across input representations was motivated by the comparatively weak effects of this factor on the correlation between audio descriptors. Figure 4 reports an agglomerative clustering model of D (median linkage; cophenetic correlation between input distances and hierarchical-clustering distances = 0.72) and a three-dimensional metric MDS model of the between-descriptor distances (SAS Institute Inc., 2010, proportion of explained variance = 0.78). It should be emphasized that although further analyses focused on the hierarchical clustering model, the MDS model is reported here for the sake of providing the reader with a complementary and easy-to-grasp representation of the raw matrix of between-descriptor correlations.

The goal of the final analysis was to estimate approximately the number of statistically independent groups of audio descriptors based on the hierarchical clustering model of the between-descriptor distances (see Fig. 4). The estimation process relied on the computation of four indices describing the quality of the clustering for each of the levels of the hierarchical model. The first clustering indices were: the point-biserial correlation, the gamma index (Milligan, 1981), and the differential clustering height, which measures the difference between the cophenetic distances of descriptors merged into N and N + 1 clusters (see Fig. 5). We computed one additional clustering index to compare the stability of the main clustering solution across the following input representations: STFTlin, STFTpow, ERBfft, ERBgam, and Harmonic. To this purpose, we computed five additional clustering solutions by choosing between each of the five input representations in turn [the modeled D matrix was computed by using the same strategy as for analysis (1), this time considering both descriptive statistics t for the same matrix]. For each of the possible numbers of clusters, we used the adjusted Rand index (Hubert and Arabie, 1985) to measure the agreement between each of the five clustering solutions on the one hand and the main clustering solution on the other. The adjusted Rand index takes values of zero and one for chance-level and perfect between-partition agreement, respectively. The bottom panel of Fig. 5 shows the median of the adjusted Rand indices across the five representation-specific clustering solutions. For all these indices, high values indicate optimal clustering partitions. It should be noted (1) that the global peak of clustering indices often favor trivial solutions (e.g., in this case, the gamma index appears to favor the trivial solution with one cluster for each descriptor) and (2) that different clustering indices might give different recommendations for an optimal partition.

For these reasons, it is recommended to estimate the optimal partition by inspecting the agreement between the indications from the local peaks for the clustering indices (Gordon, 1999). Overall, a low agreement emerged between the various clustering indices. Nonetheless, the 10-cluster solution appeared to be locally favored by three of the four indices: point biserial correlation, differential height, and adjusted Rand index. In the following, we take the ten-cluster partition as a working estimate of the optimal number of clusters of audio descriptors.

V. DISCUSSION

We analyzed the intercorrelation among audio descriptors within a large database of musical instrument tones. Several of the time-varying audio descriptors in the Timbre Toolbox can be extracted based on different variants of a spectrotemporal representation (STFT-mag, STFTpow, ERBfft, and ERBgam). A first analysis thus aimed at comparing the extent to which the structure of the intercorrelations among audio descriptors are differentially affected by a change in basic representation, or by a change in the statistical operator adopted to summarize the time-varying audio descriptors (median vs. interquartile range). The structure of the intercorrelations appeared to be weakly affected by a change in the basic input representation. For example, despite the small differences between the ERBfft and ERBgam representations, descriptors computed from the ERBfft representation that were very strongly intercorrelated were also strongly intercorrelated when computed from the ERBgam representation. For this reason, we conclude that they both give similar results and the ERBfft representation is preferable because of its greater computational efficiency and because the temporal response of all frequency channels is identical.

To the contrary, the change in statistical operator adopted to summarize the time-varying descriptors appeared to have a very strong influence on the structure of the correlation between descriptors, i.e., variations in median values across descriptors are quite different from variations in interquartile range values. For this reason, it is suggested that future studies on musical timbre take into account multiple statistics that capture different properties of the time-varying audio descriptors. In this analysis, we focused on two highly robust statistics: median and interquartile range. We did so in order to provide outlier-resistant results that would be more likely to be replicated by independent studies. Strong influences of a choice of the statistic on the between-descriptor correlations are also likely when focusing on other operators such as the mean, standard deviation or range of variation of the time-varying audio descriptor. In summary, the results of this analysis suggest that future studies on the acoustical modeling of musical sounds should focus their attention on the statistical characterization of the time-varying audio descriptors, rather than on the fine-tuning of

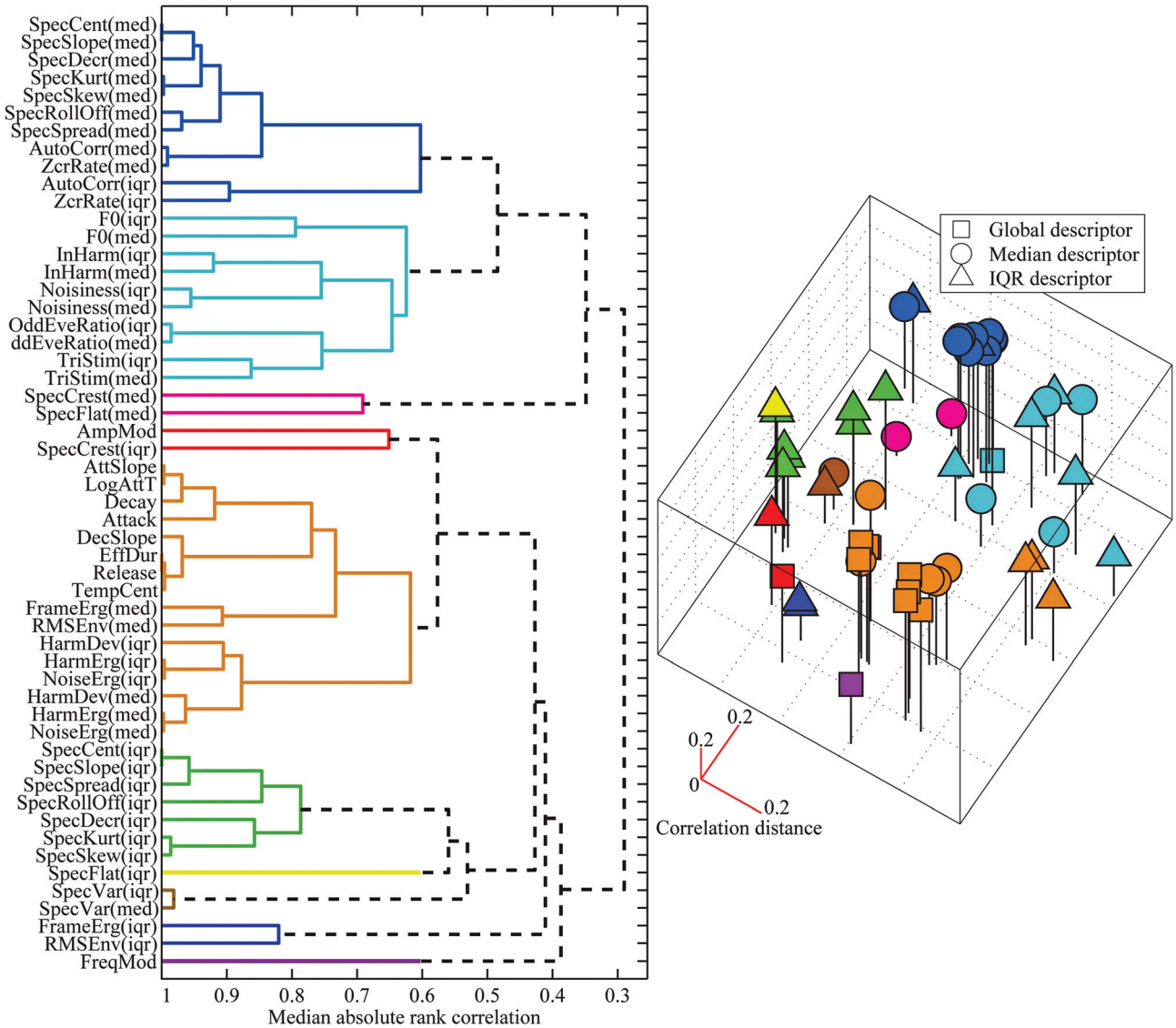

FIG. 4. Structure of similarities among audio descriptors. (Left) Hierarchical cluster analysis of the correlations among audio descriptors (med = median; iqr = interquartile range). Different colors are used to highlight different clusters of descriptors. (Right) Three-dimensional metric MDS of the between-descriptor correlations. The distance between descriptors in the MDS representation approximates a correlation distance equal to one minus the absolute correlation. In order to aid the interpretation of the MDS representation in terms of correlation distances, the MDS figure is complemented with a grid arbitrarily spaced at a correlation distance of 0.2. Within this representation, when two descriptors are separated by one grid distance their absolute correlation equals 0.8. The color codes for descriptors in the MDS representation correspond to the color codes used in the hierarchical clustering representation. See Table I for the names of audio descriptors.

the characteristics of the spectrotemporal representations from which the descriptors are extracted.

Subsequent analyses aimed at analyzing the structure of the between-descriptor correlations in greater detail. Correlation information was collapsed across spectrotemporal input representations because the initial analyses suggested a weak effect of this analysis parameter. The correlational measure of the distance between descriptors was modeled both as a hierarchical clustering tree and as a three-dimensional multidimensional scaling space. Together, these distance models are meant as a compact and easily understandable presentation of the structure of the raw correlation matrix and can be used to inspect in detail the informational redundancy of the audio descriptors implemented in the Timbre Toolbox.

A final analysis estimated the number of groups of highly intercorrelated descriptors. The goals of this analysis were to (1) assess whether the Timbre Toolbox is capable of accounting for the dimensional richness of real musical sounds, i.e., whether it is a multidimensional measurement instrument; and (2) provide the user of the Timbre Toolbox with a set of guidelines for selecting among the numerous descriptors it implements. This analysis was carried out by focusing on the hierarchical clustering model of the correlational distance between descriptors. Based on the inspection of various internal measures of clustering, we estimated ten groups of descriptors that are relatively independent from an informational point of view. As such, the Timbre Toolbox appears to be a dimensionally rich instrument for the acoustical analysis of musical sounds. Various interesting aspects emerged from the inspection of the estimated ten clusters of descriptors. Firstly, two independent clusters emerged that overall group together spectral

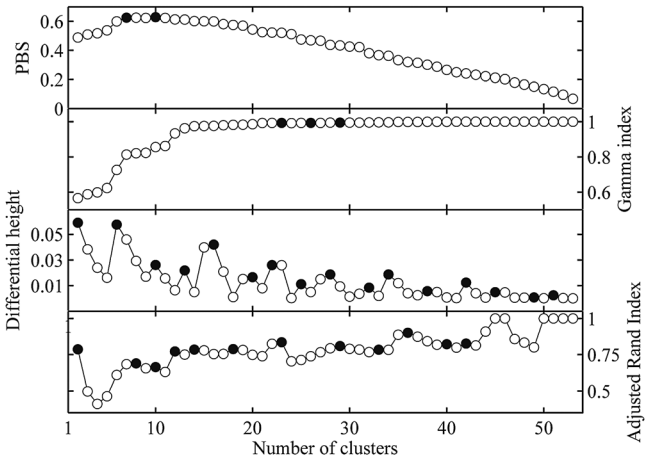

FIG. 5. Clustering indices used to estimate the number of groups of independent audio descriptors. The top three panels include point-biserial correlation (PBS), gamma index and differential clustering height. The bottom panel shows the median of the adjusted Rand index between the main clustering solution and the clustering solution obtained when choosing one of the five different input representations. Within this context, the adjusted Rand index can be interpreted as measuring the extent to which the partitioning of descriptors into clusters is affected by a change in the input representation, where high values indicate a high stability across input representations. For all indices, high values indicate better clustering solutions. Black symbols highlight the local peaks for each of the indices.

descriptors computed using the same time-varying operator. As such, one cluster mostly contained median descriptors for spectral time-varying descriptors [e.g., Centroid (med) and Kurtosis (med)], whereas another cluster included interquartile range spectral descriptors [e.g., Centroid (iqr) and Kurtosis (iqr)]. A third large cluster included the vast majority of the temporal descriptors (e.g., log-attack-time) and the energetic descriptors [e.g., NoiseErg (iqr) and TotErg (iqr)]. A final large cluster included descriptors measuring mostly signal periodicity or lack thereof (e.g., F0 and noisiness). It is interesting to note that both the temporal/energetic cluster and the periodicity cluster included descriptors computed by applying the median and interquartile range statistics to the same time-varying descriptor. As such, the correlation of these descriptors with the other audio descriptors does not appear to be strongly affected by the particular choice of the statistic summarizing the time-varying descriptor, suggesting a covariation of central tendency and variability measures for these descriptors. The same result emerged from the two-descriptor cluster including SpecVar (med) and SpecVar (iqr). As we have already discussed, this was not the case for many spectral descriptors. Very small distances can also be observed between SpecCent/SpecSlope and SpecKurt/SpecSkew and /LAT, EffDur/TempCent, and HarmErg/NoiseErg/Erg. Indeed, these are all variations of the same computational concepts. We can see that SpecVar (med and iqr) is a specific cluster, which makes sense given that it is the only spectro-temporal descriptor. The autocorrelation coefficients are related to a description of the spectral shape, which makes sense considering that xcorr is the inverse FFT of the power spectrum. And Env(iqr) and FrameErg(iqr) are fairly similar, because they only differ in low-pass filtering. The position of HarmDev is difficult to explain, however. Focusing on the issue of selecting among audio descriptors, the analysis of the clustering model suggests that future studies of musical sounds should consider at the very least: (1) one measure of the central tendency of time-varying spectral descriptors; (2) one measure of the temporal dispersion of time-varying spectral descriptors; (3) one descriptor of the energetic content of the sound signals and of the temporal envelope of energy; (4) one descriptor of the periodicity (e.g., F0 or noisiness). Notably, this minimum requirement characterizes few or none of the previous studies on the perception of musical timbre.

Focusing on the problem of estimating the number of clusters, it is important to emphasize that the 10-cluster estimates reported here should be considered as a working estimate because the goal of this analysis was not to solve exactly the problem of the optimal number of clusters. Future users of the Timbre Toolbox are thus advised to follow a method similar to that presented in this manuscript to solve this problem for custom databases of sound signals. To this purpose, a few methodological considerations are in order. For a variety of reasons (e.g., stability of parameter estimates in mathematical/statistical models), the goal of the investigators might be to focus on descriptors of musical timbre that are independent from the statistical point of view. They might either pick one descriptor from each of the clusters or reduce all the descriptors within one cluster to one single variable (e.g., by a principal-component analysis as in Giordano *et al.,* 2010). Importantly, as the number of clusters chosen by the investigator grows, descriptors become more and more correlated with each other, i.e., the differences between descriptors within the same cluster and the differences between clusters of descriptors decrease. For this reason, music-information-retrieval or machine-classification studies of musical timbres will capitalize on rather small differences in the information content of audio descriptors when their estimate of the number of clusters grows towards the number of audio descriptors. This fact is particularly important for studies on the perception of complex sounds because of the noise necessarily present in behavioral data (e.g., trial-to-trial stochastic variations in the responses of experimental participants and variability among participants). Indeed, as the estimate of the number of clusters grow, the acoustical models of behavioral data will be more and more likely to capitalize on the noise-like portions of the variance of the behavioral data rather than on those noise-free components that allow one to infer the architecture of perceptual and cognitive processes. For these reasons, the investigator should be particularly wary of choosing a very large number of clusters if this means separating moderately to highly correlated descriptors.

Finally, it is important to clarify the extent to which the analysis results presented in this study can be generalized to different sets of sound signals. The tested database of musical signals was highly heterogeneous: it comprised large variations in pitch, dynamics, duration and a large number of exemplars from each of the main families of Western musical instruments (e.g., idiophones, chordophones, electrophones). For this reason, the current analyses are likely representative of those obtained with sets of musical sounds

that comprise variations across these factors. This conclusion is strengthened by the robust nature of the adopted statistical framework (e.g., robust correlations, robust measures of central tendency), which allows for inferences that are representative of the vast majority of the datapoints and are independent of outliers. The very general approach presented in this study is a necessary first step for characterizing the rich acoustical structure of Western musical sounds and can be taken as a reference against which to test in detail the effects of various musical dimensions (e.g., pitch), the acoustical structure of non-Western musical instruments or the acoustical structure of the sounds from one single musical instrument.

VI. CONCLUSION

We have described a novel computational toolbox for extracting audio descriptors of sound signals, the Timbre Toolbox.

This toolbox extracts a large number of audio descriptors previously scattered throughout the literature on speech analysis, perception of musical timbres, sound classification and music-information retrieval. In an attempt to systematize the methods used for computing audio descriptors in the literature, we have proposed to organize the system as a set of operators (e.g., the one used to compute the spectral centroid) applied to a set of input signal representations (e.g., the linear amplitude or power DFT, the amplitude of Harmonic partials, or the outputs of ERB filters). Audio descriptors are also organized considering the temporal extent they represent. We classed them into global and time-varying descriptors. The Timbre Toolbox implements several descriptive statistics for summarizing the information about the sound signal contained in a time-varying descriptor with a single number. We propose the use of the median and interquartile range as alternatives to the mean and standard deviation because they are not sensitive to outliers present in the time-varying sequences of the audio features.

We used the Timbre Toolbox to analyze a large database of musical sounds. The overall goal of this analysis was to estimate the informational redundancy concerning the sound signals across the available sound descriptors, thus providing a useful set of audio descriptors for empirical research. We focused on measures of between-descriptor correlations as quantifiers of information redundancy. We observed that the structure of the correlations among audio descriptors is relatively robust to changes in input representation but largely affected by changes in the descriptive statistics for time-varying descriptors. From a practical point of view, this result suggests that the researcher is relatively free to choose among input representations based either on computational considerations (e.g., whichever is less demanding) or based on how accurately they model the transformations of the sound signal that take place in the auditory system. We also observed that the audio descriptors can be grouped into ten classes that are largely independent from the statistical point of view (between-group absolute correlations < 0.6). Based on this result, we conclude that the Timbre Toolbox provides informationally rich descriptions of the sound signals. The largest of these groups included (1) descriptors quantifying the central tendency of time-varying spectral properties; (2) descriptors quantifying the temporal variability of time-varying spectral properties; (3) descriptors quantifying global energetic properties and descriptors for the properties of the energy envelope; and (4) descriptors of the periodicity of the sound signal. From a practical point of view, these results suggest that multiple descriptive statistics of the time-varying descriptors should be taken into consideration because of their ability to capture different aspects of the sound signals. Although characterizing sound signals with the highest possible number of descriptors will surely maximize the amount of information extracted concerning the sound signal, it is important that researchers carry out a principled selection among descriptors based on the quantification of their informational overlap (e.g., correlational analysis) and based on considerations of the reliability of behavioral data in comparing the samples of a given sound set. The same principles can guide the process of merging informationally similar descriptors using data-reduction techniques (e.g., Giordano *et al.,* 2010).

ACKNOWLEDGMENTS

This research was supported by European ESPRIT 28793 Project Cuidad, European I.S.T. Project CUIDADO, French PRIAMM Project ECRINS, and French Oseo Project QUAERO grants to Geoffroy Peeters at Ircam, and grants from the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chair program to S.M. The authors would like to recognize the years of fruitful exchange among colleagues within and between the Music Perception and Cognition team and the Analysis/Synthesis team at Ircam. This whole project owes a great debt to Jochen Krimphoff whose initial research in Krimphoff (1993) led the way, and to Corey Kereliuk for help in programming the Timbre Toolbox. The authors wish to thank Alain de Cheveigné for sharing the code that implements the ERB filters. Part of the code in the toolbox is derived from Malcom Slaney's Auditory Toolbox (Slaney, 1998).

[1]In the music-information retrieval field, what we call "descriptor" is more often referred to as "feature." However, a distinction is made in psychology between dimensions and features. Dimensions are continuous and features are discrete. The word descriptor is agnostic as to whether the thing being described is a continuous dimension or a discrete feature, although in our case all of the descriptors take values from a continuous dimension.

[2]Indeed the presence of recording noise or breath/bow noise at the beginning of the sound makes the estimation of t_{st} fail. Further, some instruments, such as the trumpet, have a continuously increasing envelope, which makes the estimation of t_{end} fail.

Brown, J. (**1998**). "Musical instrument identification using autocorrelation coefficients," in *Proc. Intern. Symposium on Musical Acoustics* (Leavenworth, Washington), pp. 291–295.

Brown, J., Houix, O., and McAdams, S. (**2001**). "Feature dependence in the automatic identification of musical woodwind instruments," J. Acoust. Soc. Am. **109**, 1064–1072.

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

Camacho, A., and Harris, J. (**2008**). "A sawtooth waveform inspired pitch estimator for speech and music," J. Acoust. Soc. Am. **124**, 1638–1652.

de Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930.
Fujinaga, I. (**1998**). "Machine recognition of timbre using steady-state tone of acoustical musical instruments," in *Proc. of Int. Computer Music Conference* (University of Michigan, Ann Arbor, MI), pp. 207–210.
Fujinaga, I., and MacMillan, K. (**2000**). "Realtime recognition of orchestral instruments," in *Proc. of Int. Computer Music Conference* (Berlin, Germany), pp. 241–243.
Giordano, B. L., Rocchesso, D., and McAdams, S. (**2010**). "Integration of acoustical information in the perception of impacted sound sources: The role of information accuracy and exploitability," J. Exp. Psychol. **36**, 462–479.
Gordon, J. (**1987**). "The perceptual attack time of musical tones," J. Acoust. Soc. Am. **82**, 88–105.
Gordon, A. D. (**1999**). *Classification*, 2nd ed. (Chapman & Hall/CRC, Boca Raton, FL), pp. 1–256.
Grey, J. (**1977**). "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. **61**, 1270–1277.
Grey, J., and Gordon, J. (**1978**). "Perceptual effects of spectral modifications on musical timbres," J. Acoust. Soc. Am. **63**, 1493–1500.
Herrera, P., Amatriain, X., Batlle, E., and Serra, X. (**2000**). "Towards instrument segmentation for music content description: A critical review of instrument classification techniques," in *International Symposium on Music Information Retrieval (ISMIR 2000)*, Plymouth, MA (University of Massachusetts, Amherst, MA), pp. 23–25.
Hubert, L. J., and Arabie, P. (**1985**). "Comparing partitions," J. Classif. **2**, 193–218.
ISO/IEC (**2002**). "MPEG-7: Information Technology – Multimedia Content Description Interface - Part 4: Audio (ISO/IEC FDIS 15938-4:2002)."
Iverson, P., and Krumhansl, C. (**1993**). "Isolating the dynamic attributes of musical timbre," J. Acoust. Soc. Am. **94**, 2595–2603.
Johnston, J. (**1988**) "Transform coding of audio signals using perceptual noise criteria," IEEE J. Sel. Areas Commun. **6**(2), 314–323.
Kendall, R., Carterette, E., and Hajda, J. (**1999**). "Perceptual and acoustical features of natural and synthetic orchestral instrument tones," Music Percept. **16**(3), 327–364.
Krimphoff, J. (**1993**). "Analyse acoustique et perception du timbre [Acoustic analysis and perception of timbre]," Master's thesis (Université du Maine, Lemans, France).
Krimphoff, J., McAdams, S., and Winsberg, S. (**1994**). "Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique (Characterization of the timbre of complex sounds. II Acoustical analysis and psychophysical quantification)," J. Phys. **4**, 625–628.
Krumhansl, C. L. (**1989**). "Why is musical timbre so hard to understand?," in *Structure and Perception of Electroacoustic Sound and Music*, edited by S. Nielzén Olsson (Excerpta Medica, Amsterdam, Netherlands), pp. 43–51.
Lakatos, S. (**2000**). "A common perceptual space for harmonic and percussive timbres," Percept. Psychophys. **62**(7), 1426–1439.
Maher, R., and Beauchamp, J. (**1994**). "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," J. Acoust. Soc. Am. **95**(4), 2254–2263.
Marozeau, J., and de Cheveigné, A. (**2007**). "The effect of fundamental frequency on the brightness dimension of timbre," J. Acoust. Soc. Am. **121**(1), 383–387.
Marozeau, J., de Cheveigné, A., McAdams, S., and Winsberg, S. (**2003**). "The dependency of timbre on fundamental frequency," J. Acoust. Soc. Am. **114**(5), 2946–2957.
Martin, K., Scheirer, E., and Vercoe, B. (**1998**). "Music content analysis through models of audition," in *Proc. ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications*, Bristol, UK (ACM, New York).
McAdams, S. (**1993**). "Recognition of sound sources and events," in *Thinking in Sound: The Cognitive Psychology of Human Audition*, edited by S. McAdams and E. Bigand (Oxford University Press, Oxford, UK), pp. 146–198.
McAdams, S., Windsberg, S., Donnadieu,S., DeSoete, G., and Krimphoff, J. (**1995**). "Perceptual scaling of synthesized musical timbres: Common dimensions specificities and latent subject classes," Psychol. Res. **58**, 177–192.
McAulay, R., and Quatieri, T. (**1986**). "Speech analysis/synthesis based on a sinusoidal representation," IEEE Trans. Acoust. Speech Signal Process. **34**(4), 744–754.
Miller, J., and Carterette, E. (**1975**). "Perceptual space for musical structures," J. Acoust. Soc. Am. **58**, 711–720.
Milligan, G. W. (**1981**). "A Monte Carlo study in thirty internal criterion measures for cluster analysis," Psychometrika **46**, 187–199.
Misdariis, N., Smith, B., Presssnitzer, D., Susini, P., and McAdams, S. (**1998**). "Validation of a multidimensional distance model for perceptual dissimilarities among musical timbres," J. Acoust. Soc. Am. **103**, 3005–3006.
Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.
Opolko, F., and Wapnick, J. (**2006**). McGill University Master Samples [DVD set] (McGill University, Montréal, Québec, Canada).
Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (**1992**). "Complex sounds and auditory images," Aud. Physiol. Percept. **83**, 429–446.
Peeters, G. (**2004**). "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," CUIDADO IST Project Report (IRCAM, Paris), pp. 1–25.
Peeters, G., McAdams, S., and Herrera, P. (**2000**). "Instrument sound description in the context of MPEG-7," in *Proc. of Int. Computer Music Conference*, Berlin, Germany (ICMA, San Francisco).
Plomp, R. (**1970**). "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg (Sijthoff, Leiden), pp. 397–414.
Pollard, H., and Jansson, E. (**1982**). "A tristumulus method for the specification of musical timbre," Acustica **51**, 162–171.
Rioux, V., McAdams, S., Susini, P., and Peeters, G. (**2002**). "Wp2.1.5. psycho-acoustic timbre descriptors," Cuidado Report (IRCAM, Paris).
SAS Institute Inc. (**2010**). SAS/Stat 9.22 User's Guide (SAS Institute Inc., Cary, NC).
Scheirer, E., and Slaney, M. (**1997**). "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. of IEEE Int. Conference on Acoustic Speech and Signal Processing*, Munich, Germany (IEEE Computer Society Press, Los Alamitos, CA), Vol. 2, pp. 1331–1334.
Serra, X., and Smith III, J. (**1990**). "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," Comput. Music J. **14**, 12–24.
Slaney, M. (**1998**). "Auditory Toolbox, Version 2, Technical Report No: 1998-010" (Interval Research Corporation).
Tindale, A., Kapur, A., Tzanetakis, G., and Fujinaga, I. (**2004**). "Retrieval of percussion gestures using timbre classification techniques," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2004)*, Barcelona, Spain (Audiovisual Institute Pompeu Fabra University), pp. 541–544.
Wedin, L., and Goude, G. (**1972**). "Dimension analysis of the perception of instrumental timbre," Scand. J. Psychol. **13**, 228–240.
Wessel, D. (**1979**). "Timbre space as a musical control structure," Comput. Music J. **3**, 45–52.
Wessel, D. L. (**1973**). "Psychoacoustics and music: A report from Michigan State University," PACE: Bull. Comput. Arts Soc. **30**, 1–2.
Wright, M. (**2008**). "The shape of an instant: Measuring and modeling perceptual attack time with probability density functions," Ph.d. thesis, Stanford University, Palo Alto, CA.
Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," J. Acoust. Soc. Am. **33**, 248.