Audio Descriptive Synthesis "AUDESSY"

Eddy Savvas Kazazis

Institute of Sonology

Royal Conservatory in The Hague

Master's Thesis

2014 May

Abstract

This thesis examines the viability of audio descriptors within a synthesis context. It provides insight into acoustical modeling based on verbal descriptions by quantifying the relationships between verbal attributes of timbre and a set of audio descriptors. Various predictive models of verbal attribute magnitude estimation (VAME) are also tested. The results show that is possible to create, classify and order sounds according to a verbal description. Finally, audio descriptive synthesis (AUDESSY) is introduced. This technique offers the possibility to synthesize and modulate sounds according to sonic morphologies, which are revealed by audio descriptors.

Keywords: timbre, timbre space, audio descriptors, sonic morphology, perception, optimization, synthesis, analysis, dimensional reduction, partial least squares regression.

To *Paul Berg.*

"We long ago quit talking about "happy melodies" and "pungent harmonies" in favor of contextual musical analysis of developing musical structures of, primarily, pitch and rhythm; and I would hope that we could soon find whatever further excuse we still need to quit talking about "mellow timbres" and "edgy timbres," and "timbres" altogether, in favor of contextual musical analysis of developing structures of vibrato, tremolo, spectral transformation, and all those various dimensions of sound which need no longer languish as inmates of some metaphor."

J. K. Randall: Three lectures to scientists.

(Randall, 1967)

Acknowledgements

I would like to thank my mentor Paul Berg. Stephen McAdams for accepting me in his team and co-supervising part of this thesis. Kees Tazelaar for his support and his efforts that led to such collaboration between the Institute of Sonology and McGill University. Johan van Kreij, Peter Pabon and Joel Ryan for their comments, enthusiasm and fruitful discussions. Kristoffer Jensen for a very warm discussion during a cold winter day in Copenhagen. Researchers and future colleagues at the Music Technology Area of McGill: Bennett Smith, Sven-Amin Lembke, Kai Siedenburg, Cecilia Taher and Charalambos Saitis. Asterios Zacharakis for providing us the results of his study. Svetlana Jovanovic. My friend Pavlos Kranas. My family.

Contents

List of Figures

LIST OF FIGURES

List of Tables

Chapter 1

Introduction

The main motivation for this work arises from the author's general interest in *timbre*, and the notion that *sound*, is a structured entity that can be apprehended through a compact and qualitative representation.

1.1 Timbre: A Word versus a Phenomenon

Paul Berg gives a rather poetic definition of timbre:

> "Timbre is Magic." (Berg, 2012)

We can either add more mystery into the subject by quoting Denis Smalley's conclusion, drawn from his contradictory article *Defining Timbre - Refining Timbre*:

> "Timbre is dead. Long live timbre." (Smalley, 1994)

Or, we can demystify what timbre is about by paying too much attention to the negative definition provided by the American Standards Association (ASA, 1960). Al Bregman in *Auditory Scene Analysis* nicely puts that definition in question:

> "The problem with timbre is that it is the name for an ill-defined waste-basket category. Here is the much-quoted definition of timbre given by the American Standards Association: 'that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.' This is, of course, no definition at all." (Bregman, 2001)

Indeed, the *word* timbre has a catch. Robert Erickson in *Sound Structure in Music* cites the translator of Hermann Helmholtz's *On the Sensations of Tone*, Alexander Ellis:

> " ... Timbre, properly a kettledrum, then a helmet, then the coat of arms surmounted with a helmet, then the official stamp bearing that coat of arms (now used in France for a postage label), and then the mark which declared a thing to be what it pretends to be ..." (Erickson, 1975)

Stephen McAdams comments further the vagueness of the *word* and points out some isles that timbre leaves its marks on:

> "Timbre is a misleadingly simple and exceedingly vague word encompassing a very complex set of auditory attributes, as well as a plethora of intricate psychological and musical issues." (McAdams, 2013)

As a conclusion, we might say that timbre is a multidimensional phenomenon bounded by context, listening strategies and listening abilities.

1.2 A Qualitative Representation of Sound

In the present thesis we examine the relationships between verbal attributes of timbre and their acoustic correlates, and we attempt to determine sonic morphologies as these arise by a purely descriptive model. If the analysis model is too general, it will be incapable to reveal any morphology at all. On the other hand, if it is too specific, the essence of what it is assumed to describe might be lost due to highly redundant information.

A nice compromise between these two extremes could be made if we build our analysis model by extracting some carefully chosen audio features, which we shall call "audio descriptors"[1]. Once the morphologies are determined, we attempt to synthesize sounds that encapsulate the desired characteristics while at the same time leaving room for further artistic exploration. This can be seen as an analysis by synthesis approach (Risset, 1991) which allows us to: synthesize sounds starting from a description of

[1]Essentially, "audio descriptors" are acoustic parameters that correlate with a perceptual dimension though in the present thesis they will refer to audio features.

their physical structure; model and synthesize sounds, based on perceptual dimensions; morph between sounds; create generic sound templates.

"Audio descriptors", as the name suggests, don't define, rather describe sound and hopefully this is a well-known concept to composers. A score is condemned to describe what it refers to since the elements that define music emerge from the actual performance.

1.3 Framework

Initially, we create sound profiles by performing an *audio descriptive analysis* on a gamut of sounds. These "templates" will be used to demonstrate some concepts throughout this thesis and will serve as general guidelines for the synthesis process.

1.3.1 Analysis

The analysis process is quite straightforward and most of the times will be performed so that its results can be directly used as synthesis-parameters. First of all, we need to choose an *input representation* of the sound being analyzed, according to the audio features that are to be extracted. We can choose one or more from the following representations: *the temporal energy envelope*; *the short-time Fourier transform*; *a sinusoidal model*; and a more "strict" *harmonic representation*.

Afterwards, we specify the set of audio descriptors that we intend to use by considering the appropriateness of each descriptor with respect to the synthesis scheme[2], and by taking into account the fact that audio descriptors are often (and sometimes highly) inter-correlated. Finally, we apply various operators to the input-representation of the signal to derive the audio descriptors.

1.3.2 Synthesis

The synthesis scheme requires a *source* sound that will be represented by a sinusoidal model and eventually, all operations will act upon this representation. The source can be any waveform (sampled or directly specified in the sinusoidal model), which will be transformed according to a *target-morphology* as dictated by the audio descriptors.

[2]For example, it would be odd to use the *zero-crossing rate* as a parameter in a frequency-domain synthesis algorithm.

In the next stage, we extract the audio descriptors and set their *target* values. These values can either be derived from the previous analysis stage, or they can be specified according to a preconceived sonic morphology. We force the source to adopt the target values by utilizing an *optimization* algorithm, using the audio descriptors as constraints, and as an objective function the sum of partials' amplitudes, which are obtained by the sinusoidal representation. If there is a feasible solution, the optimization will lead us to obtain *the best* sound, in a sense that it will be as close as possible to the source sound while at the same time will ensure that all of our constraints are satisfied (i.e. the audio descriptors have attained their target values). Finally, we apply the results obtained by the optimization process to the sinusoidal model and convert it back to sound by using *additive synthesis*.

1.4 Structure of the Document

Chapter 2 introduces the audio descriptors that will be used in the present thesis and gives a brief presentation of the Timbre Toolbox (Peeters, Giordano, Susini, Misdariis & McAdams, 2011), which is an analysis toolbox built in MATLAB.
Chapter 3 focuses on the perceptual saliency of spectral, temporal and spectrotemporal attributes of timbre, and the construction of *timbre spaces*.
Chapter 4 presents the conclusions from previous studies on timbre semantics, and creates a link between verbal attributes of timbre and audio descriptors, aiming to provide insight into acoustical modeling based on perceptual dimensions.
Chapter 5 explains in more depth the synthesis process and presents some plausible uses of *audio descriptive synthesis*.

Chapter 2

Audio Descriptive Analysis

Audio descriptors refer to the acoustical parameters of an audio signal, which can serve as potential physical correlates of perceptual dimensions. The formalization of these parameters over the past years has led to the development of a large set of audio descriptors, which are used in standards such as the MPEG7 (Peeters, McAdams & Herrera, 2000) and more recently in MATLAB toolboxes such as the MIRtoolbox (Lartillot & Toiviainen, 2007) and the Timbre Toolbox (Peeters et al., 2011), which will be discussed shortly.

Extracting such parameters from audio signals offers a systematic approach for deriving sonic morphologies and examining their reflections to human perception. How we gain control over these parameters will be discussed in detail in chapter 5.

In the following paragraphs we start with a brief presentation of the Timbre Toolbox. Then, we examine in relation to our methodology, the usability of input-representations from which the audio descriptors are derived. Finally, we present a formalization of audio descriptors and carry out a principled selection based on their suitability within a synthesis context.

2.1 The Timbre Toolbox

Timbre Toolbox contains a set of 32 audio descriptors that are extracted from the following input-signal representations: temporal energy envelope, short-time Fourier transform, harmonic sinusoidal components and a model of peripheral auditory processing –the Equivalent Rectangular Bandwidth (ERB) model.

These descriptors (summarized in Figure 2.1) capture temporal, energetic, spectral and spectrotemporal properties of the sound being analyzed. Temporal descriptors refer to properties such as log-attack time, decay, release and the amplitude and frequency modulation. Energetic descriptors include the harmonic-to-noise energy ratio of the signal. The spectral shape can be derived from descriptors such as the spectral centroid and higher order statistics, spectral decrease and spectral crest. Spectral variation (often called spectral flux) is the only descriptor referring to the spectrotemporal properties of the sound.

Timbre Toolbox is designed to extract audio descriptors from a single acoustic event rather from a series of events. Therefore, descriptors are divided in two categories: global descriptors, which have a single value (eg. the attack time) and time-varying descriptors, which are extracted from a frame-by-frame analysis and therefore have multiple values along the duration of the sound event. In order to have an overview of these time-varying values, descriptive statistics are used. These include the minimum or maximum values, the standard deviation, the mean and the more robust measures of central tendency and variability, expressed by the median value and interquartile ranges respectively.

2.2 Input Representations

The audio descriptive analysis is performed using the input representations presented in the next paragraphs. Audio descriptors are often inter-correlated, especially when they are applied to a limited sound-set. Peters et al. (2011) found that the inter-correlations are weakly affected between different input representations. The same is not true when applying statistical operators to time varying descriptors: a change in the operator strongly affects the structure of the inter-correlations. However, in order to summarize the behavior of time varying descriptors we are using only the median values. It should also be noted that we normalize the signal before obtaining any input representation.

2.2.1 Short-time Fourier Transform (STFT)

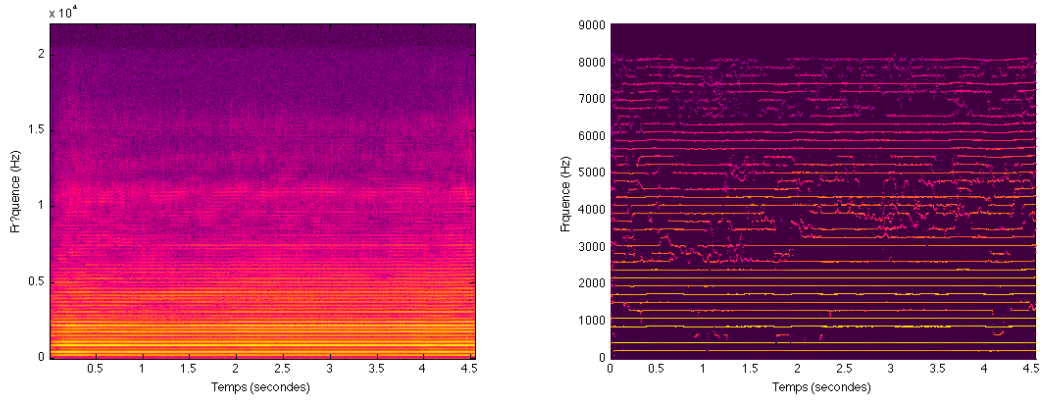
The STFT representation is obtained by using a Hamming analysis-window of 1024 points with a hop size of 256 points. The audio descriptors can then be derived from the

	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	

Figure 2.1: 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. [From Peeters et al. (2011)]

amplitude spectrum of the STFT. This representation will be used mainly for examining noisy signals that cannot be represented adequately by a sinusoidal or harmonic model.

For reconstruction purposes, sometimes it will be useful to split the signal to its sinusoidal (deterministic) and noise (stochastic) parts. We achieve this decomposition by using the *adaptive sub-band analysis and fast high resolution (HR) subspace tracking* method, which is implemented in the DESAM Toolbox (Lagrange et al., 2010). The stochastic part can then be analyzed by the STFT representation and the deterministic part, by a harmonic or sinusoidal representation.



(a) **(b)**

Figure 2.2: Signal decomposition based on the fast HR subspace tracking method: Fourier spectrogram of a violin sound (a) and its deterministic part (b).

2.2.2 Harmonic Representation

Harmonic descriptors such as the tristimulus values, the odd-to-even ratio, and inharmonicity, can only be derived from a harmonic representation. In Timbre Toolbox the input signal is analyzed using a Blackman window of 100ms with a hop size of 25ms.

Afterwards, a reference-partial is defined by estimating the fundamental frequency for each frame. The harmonic (or quasi-harmonic) partials can then be computed, such that the content and energy of the spectrum is best explained. The total number of computed partials defaults to 20 though it can be increased as much as the estimated fundamental frequency allows for.

2.2.3 Sinusoidal Representation (based on SPEAR)

When we synthesize (or resynthesize) sounds based on additive-synthesis, the sinusoidal model is the most appropriate representation for deriving audio descriptors and adds a lot of flexibility during the synthesis stages. In order to make our approach more accessible to sonologists, we derive this representation by using SPEAR (Klingbeil, 2009). Though it was originally conceived as software to aid spectral composition, it suits the needs of this thesis by proving to be a reliable tool for analysis and resynthesis.

SPEAR performs partial tracking using a variation of the *McAulay-Quatieri* technique along *linear prediction* of the partial amplitudes and frequencies, as to determine the best continuation for the sinusoidal tracks.

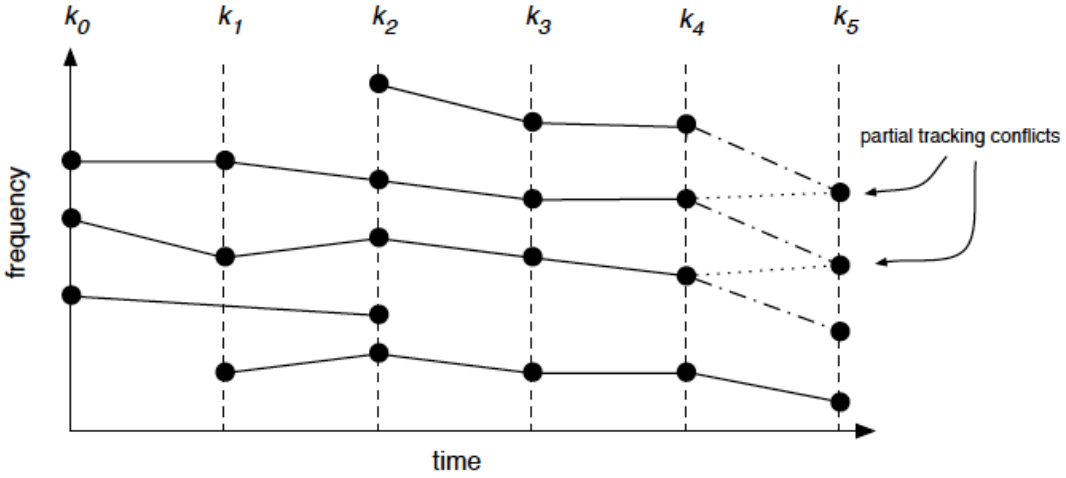


Figure 2.3: A problem that linear prediction has to solve: partial tracking conflicts due to glissandi. k_i is the frame number. [From Klingbeil (2009)]

Users can specify the length of a Blackman analysis-window as well as the amplitude thresholds above which the sinusoidal components are computed. Tracking only the perceptual significant partials eliminates redundant information and results in faster computations and easier manipulations.

The results of this analysis can be exported as a text file from which we can gather the necessary data to compute the audio descriptors. The text format is shown in Figure 2.4.

```
par-text-frame-format
point-type index frequency amplitude
partials-count <J>
frame-count <K>
frame-data
<frame-time0> <N> <index0> <freq0> <amp0> ... <indexN> <freqN> <ampN>
<frame-time1> <N> <index0> <freq0> <amp0> ... <indexN> <freqN> <ampN>
...
<frame-timeK> <N> <index0> <freq0> <amp0> ... <indexN> <freqN> <ampN>
```

Figure 2.4: The *par-text-frame-format* specification. Following the frame-data line, each line contains the breakpoints for one frame. N indicates the number of peaks in each frame. The index values connect peaks from frame to frame. Each line is separated by a newline character. [From Klingbeil (2009)]

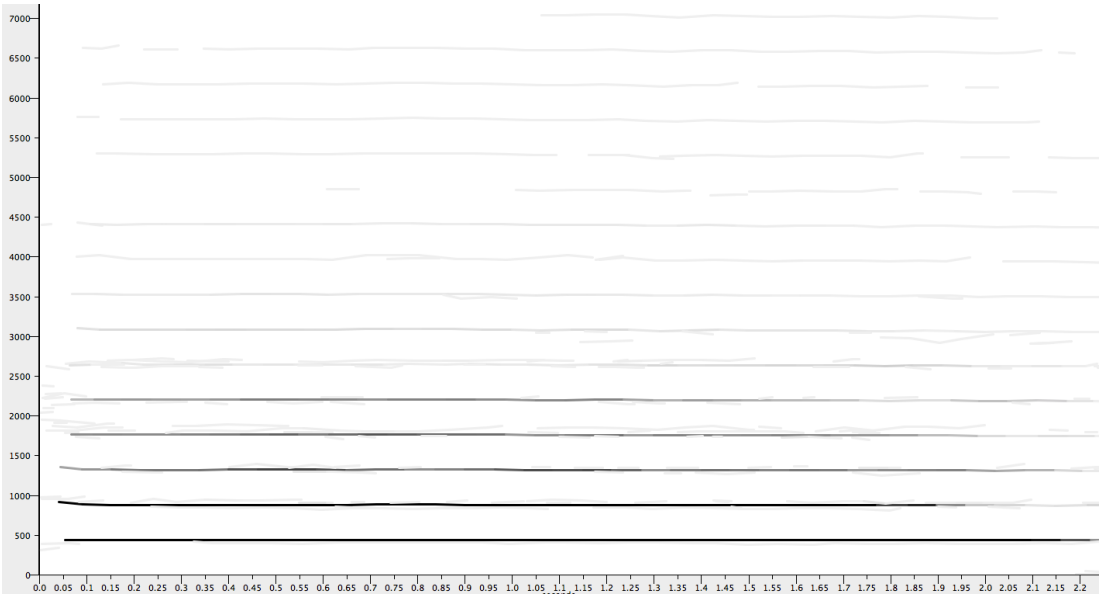


Figure 2.5: A screenshot of SPEAR. The analysis is performed on the deterministic part of a flute sound. y-axis represents frequency (in Hz). x-axis represents time (in seconds).

Sinusoidal modeling might fail to represent accurately dense polyphonic material, noisy or reverberated signals and sounds with sharp transients. However, audio descriptors are not meant to capture the fine details of the spectrum and in the present thesis, manipulating and transforming the original material is more crucial than achieving a perfect reconstruction.

2.2.4 Temporal Energy Envelope

The temporal envelope of the input-signal can be derived either from the Timbre Toolbox or the sinusoidal representation. Timbre Toolbox derives the temporal envelope from the amplitude of the analytic signal given by the Hilbert transform. In a sinusoidal representation it is derived simply by calculating the sum of partial amplitudes for each frame.

2.3 A formalization of Audio Descriptors

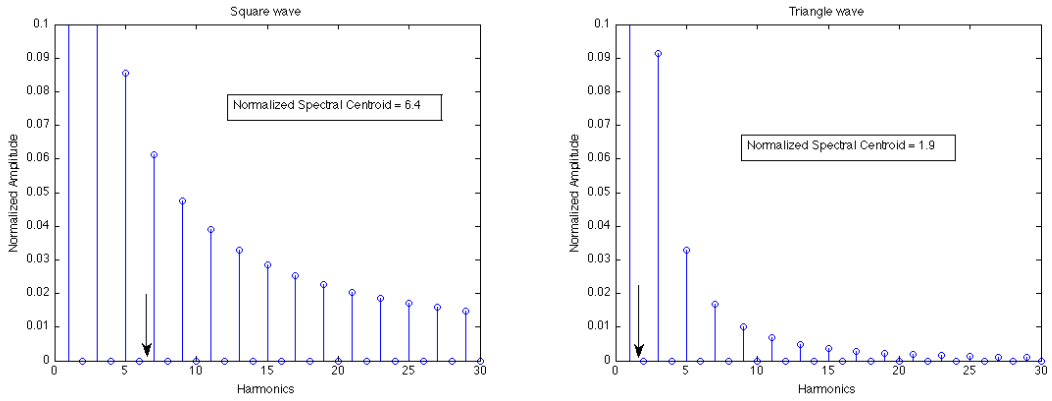
In this section, we present a formalization of audio descriptors that are drawn from the Timbre Toolbox and are relevant to the present study. Formalizations of a broader class of descriptors can be found in: Peeters (2004); Lartillot and Toiviainen (2007); Peeters et al. (2011). In the following, $f_h(t_m)$ and $a_h(t_m)$ denote the frequency and amplitude of the h^{th} STFT bin or partial at time t_m. $p_h(t_m)$ is the normalized amplitude: $p_h(t_m) = a_h(t_m)/\sum_{h=1}^{H} a_h(t_m)$, where H is the total number of bins or partials.

- **Spectral centroid** is the spectral center of gravity (Figure 2.6):

$$m_1 = \sum_{h=1}^{H} f_h p_h(t_m) \tag{2.1}$$

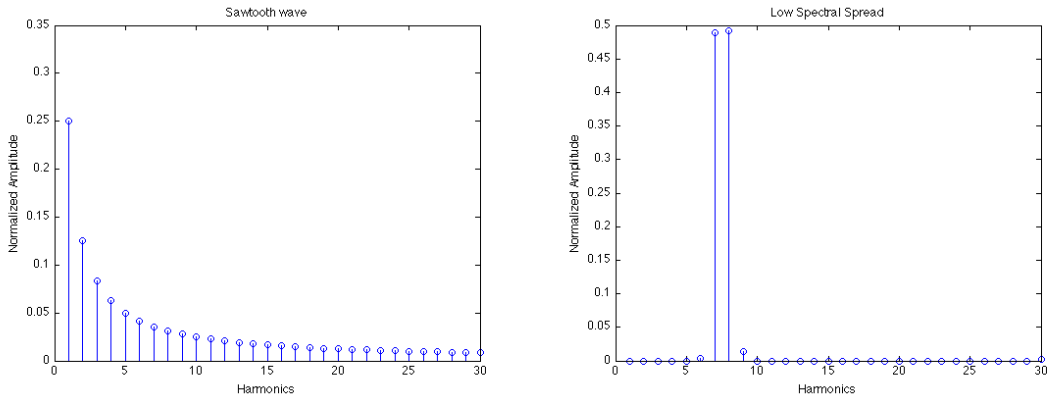
- **Spectral spread** (or spectral standard deviation) represents the spread of the spectrum around the spectral centroid (Figure 2.7):

$$m_2 = (\sum_{h=1}^{H} (f_h - m_1(t_m))^2 p_h(t_m))^{1/2} \tag{2.2}$$



(a) **(b)**

Figure 2.6: Spectral centroids: (a) has a higher spectral centroid than (b).



(a) **(b)**

Figure 2.7: Spectral spread: (a) has a higher spectral spread than (b).

- **Spectral skewness** measures the asymmetry of the spectrum around the spectral centroid. $m_3 < 0$ indicates that there is more energy at frequencies lower than the spectral centroid, $m_3 > 0$ more energy at higher frequencies and $m_3 = 0$ a symmetric distribution (Figure 2.10):

$$m_3 = (\sum_{h=1}^{H}(f_h - m_1(t_m))^3 p_h(t_m))/m_2^3 \tag{2.3}$$

- **Spectral kurtosis** measures the flatness of the spectrum around the spectral centroid:

$$m_4 = (\sum_{h=1}^{H}(f_h - m_1(t_m))^4 p_h(t_m))/m_2^4 \tag{2.4}$$

- **Spectral decrease** averages the set of slopes between frequencies f_h and f_1:

$$decrease(t_m) = \frac{1}{\sum_{h=2}^{H} a_h(t_m)} \sum_{h=2}^{H} \frac{a_h(t_m) - a_1(t_m)}{h-1} \tag{2.5}$$

- **Spectral roll-off** is the frequency $f_c(t_m)$ below which 95% of the signal energy is contained:

$$\sum_{f=0}^{f_c(t_m)} a_f^2(t_m) = 0.95 \sum_{f=0}^{sr/2} a_f^2(t_m), \qquad \text{where } sr \text{ is the sample rate.} \tag{2.6}$$

- **Tristimulus values** are three different energy ratios of the harmonics. (Figure 2.8):

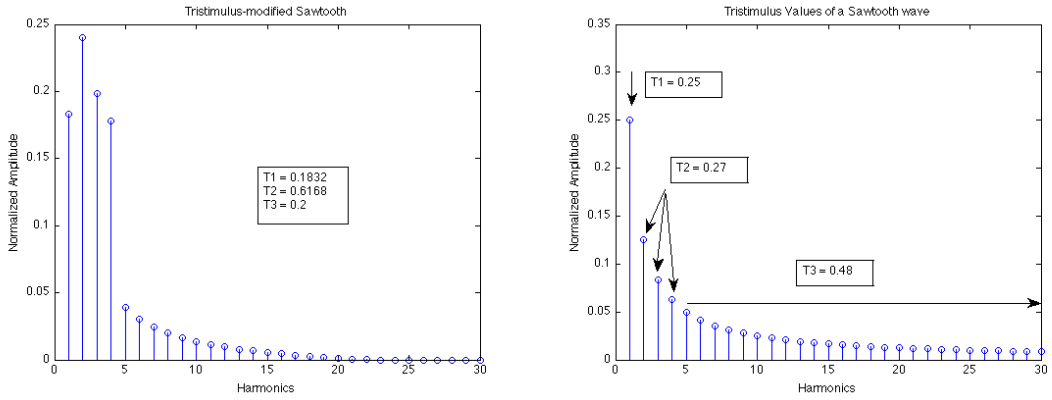
$$T1(t_m) = \frac{a_1(t_m)}{\sum_{h=1}^{H} a_h(t_m)} \tag{2.7}$$

$$T2(t_m) = \frac{a_2(t_m) + a_3(t_m) + a_4(t_m)}{\sum_{h=1}^{H} a_h(t_m)} \tag{2.8}$$

$$T3(t_m) = \frac{\sum_{h=5}^{H} a_h(t_m)}{\sum_{h=1}^{H} a_h(t_m)} \tag{2.9}$$

- **Inharmonicity** measures the deviation of partials' frequencies f_h from purely harmonic frequencies hf_0:

$$inharmonicity(t_m) = \frac{2}{f_0(t_m)} \frac{\sum_{h=1}^{H} |(f_h(t_m) - hf_0(t_m))| a_h^2(t_m)}{\sum_{h=1}^{H} a_h^2(t_m)} \tag{2.10}$$



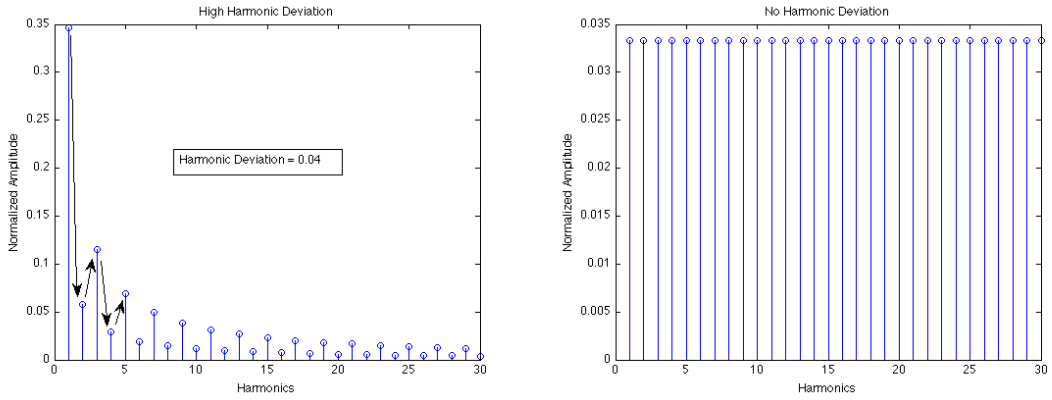
(a) **(b)**

Figure 2.8: Tristimulus values.

- **Spectral deviation** measures the deviation of partials' amplitudes from a smoothed envelope SE (Figure 2.9):

$$deviation(t_m) = \frac{1}{H}\sum_{h=1}^{H}(a_h(t_m) - SE(f_h, t_m)) \tag{2.11}$$

$$SE(f_h, t_m) = \frac{1}{3}(a_{h-1}(t_m) + a_h(t_m) + a_{h+1}(t_m)), \qquad 1 < h < H \tag{2.12}$$



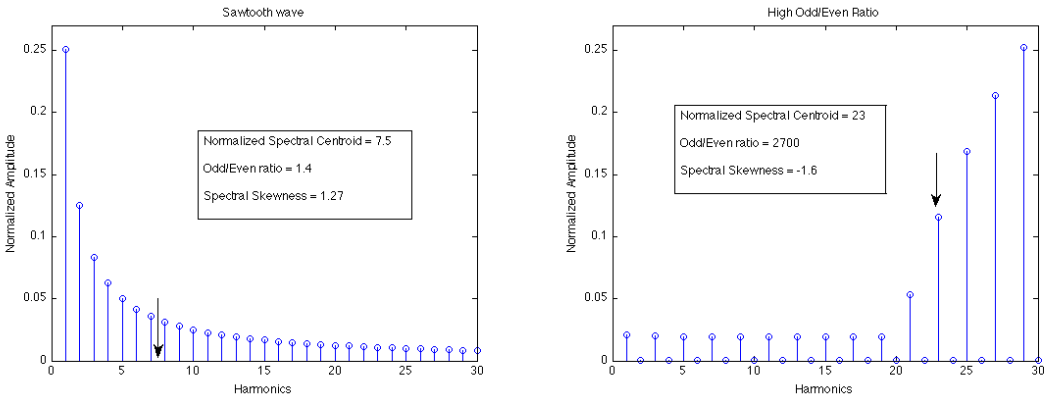
(a) **(b)**

Figure 2.9: Harmonic spectral deviation.

- **Odd-to-Even ratio** is the energy ratio of odd harmonics to even harmonics

(Figure 2.10):

$$oer(t_m) = \frac{\sum_{h=1}^{H/2} a_{2h-1}^2(t_m)}{\sum_{h=1}^{H/2} a_{2h}^2(t_m)} \tag{2.13}$$



(a) **(b)**

Figure 2.10: Two waveforms with extreme odd-to-even ratios. (a) has positive skewness and (b) has negative.

- **Spectral variation** is a measure of spectral flux. It represents how the spectrum varies over time (Figure 2.11):

$$variation(t_m, t_{m-1}) = 1 - \frac{\sum_{h=1}^{H} a_h(t_{m-1}) a_h(t_m))}{\sqrt{\sum_{h=1}^{H} a_h(t_{m-1})^2} \sqrt{\sum_{h=1}^{H} a_h(t_m)^2}} \tag{2.14}$$

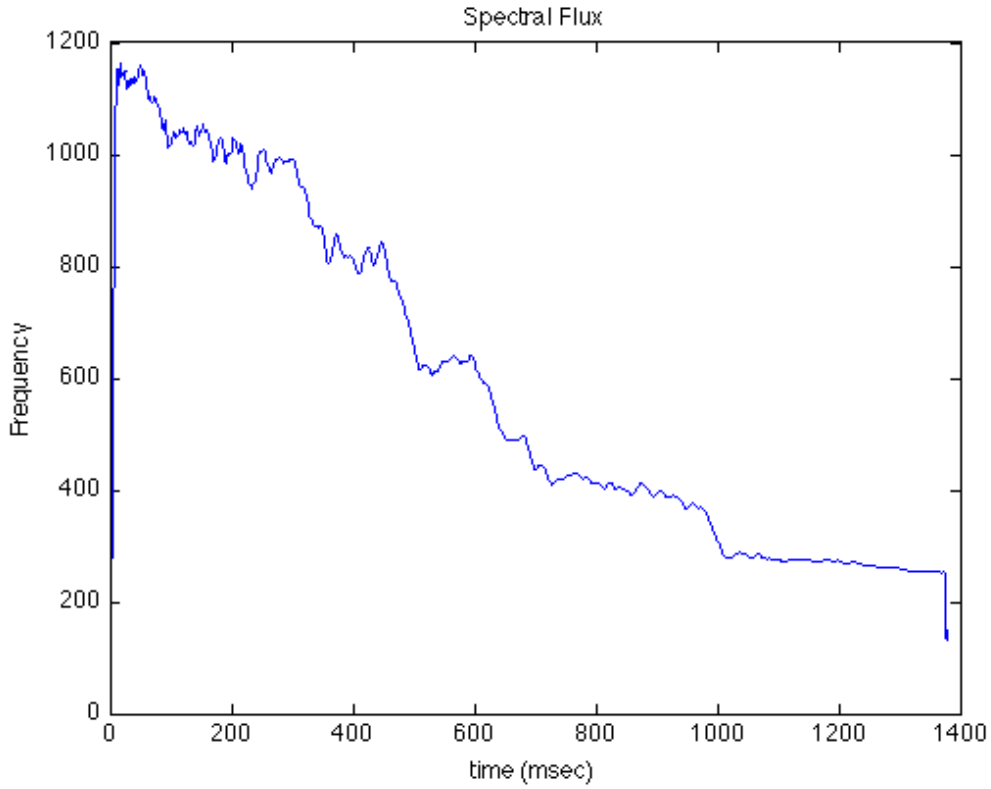


Figure 2.11: Spectral variation of an electric guitar sound.

Chapter 3

Spectral, Temporal and Spectrotemporal Attributes of Timbre

In this chapter we present a brief review of results and conclusions from past experiments that led to our current understanding of timbre and how they relate to audio descriptors. Summarizing these conclusions will help us to use audio descriptors in a synthesis context more effectively.

In general, there are two approaches used in timbre studies: in the first one, the researcher performs unidimensional studies by directly measuring presumed timbre attributes; the second approach requires no presumptions regarding the nature and number of attributes. The researcher measures relationships between stimuli and the timbre structure is uncovered by using multidimensional scaling techniques.

3.1 Unidimensional Studies

Lichte (1941) proposed that *brightness, roughness* and *fullness* are attributes that can be found in any complex sound: *brightness* was defined as a function of the midpoint of the energy distribution among partials; *roughness* was associated with the presence of high partials and their relative location (i.e. inharmonicity) in the frequency continuum; *fullness* was associated with the odd to even ratio of the partials. He also suggested that *roughness* and *brightness* could be thought of as being different functions of the

same variable, which defines the complexity of frequency ratios among partials.

Other examples of unidimensional studies can be found in early instrument-identification experiments. Berger (1964) was among the first to study the effect of temporal attributes. Listeners were asked to identify recorded wind-instrument tones when played-back: unaltered; backwards; with their attack and decay portions suppressed; and through a 480 Hz low pass filter[1]. Identification was most perturbed by the filtering process, then by attack and decay suspension and least from reverse playback.

3.2 Multidimensional Studies

In multidimensional studies listeners make paired comparisons of the sound stimuli by judging their similarity. Sounds are usually equalized in terms of pitch, loudness and perceived duration, as to shift listener's focus to a more restricted set of timbre attributes. The measurements are made on a numerical scale ranging from "identical" or "very similar" to "very dissimilar".

Multidimensional scaling (MDS) transforms the dissimilarity ratings into distances represented in a multidimensional space. As a result, perceptually similar sounds appear close together and dissimilar sounds are farther apart. The dimensionality of the MDS solution can be decided a-priori by the researcher, or determined by using a statistical criterion or a goodness-of-fit measure.

The basic MDS model (Kruskal, 1964) assumes that timbres differentiate only by the same continuous dimensions. Extended MDS models such as the EXSCAL (Winsberg & Carroll, 1989) can account for additional dimensions or distinguishing features that are specific to individual sounds among the stimuli, called "specificities". Models like INDSCAL (Miller & Carterette, 1975) and CLASCAL (McAdams, Winsberg, Donnadieu, De Soete & Krimphoff, 1995), in addition to specificities, use weights to examine how much the judgments of an individual listener rely on each dimension, or to sort listeners into different classes such as non-musicians, music-students and professionals.

The final step of the analysis is the psychophysical interpretation of the dimensions and relies heavily on the intuition of the researcher. As the number of dimensions grows the model will better explain the ratings of the listeners, but the interpretation of the

[1]Recordings were made at F4 concert pitch corresponding to approximately 349 Hz.

dimensions becomes more difficult. A relationship between the perceptual dimensions and acoustical parameters is found by computing correlations between the location of sounds on each axis, and a number of physical parameters such as spectral centroid or attack time.

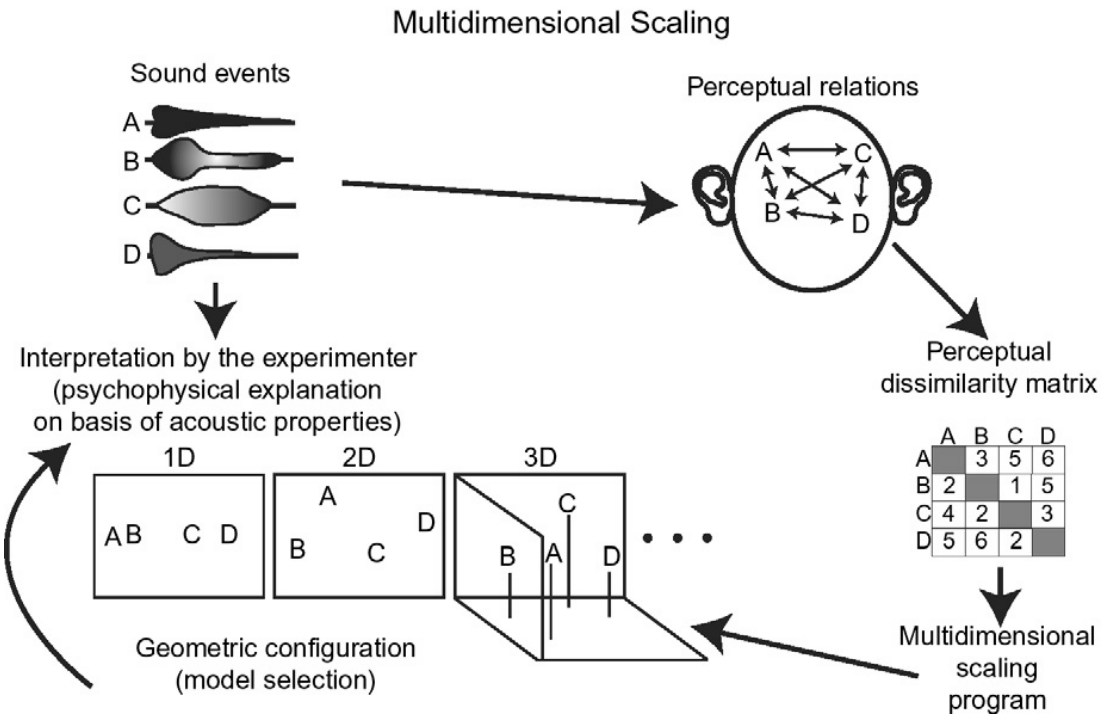


Figure 3.1: Stages in the multidimensional analysis of dissimilarity ratings of sounds differing in timbre. [From McAdams (2013)]

Plomp (1970, 1976) was among the first to use multidimensional scaling in timbre studies using Kruskal's MDSCAL model. Subjects rated the similarity of synthesized steady-state spectra derived from recorded instrument tones. The MDS solution yielded two dimensions for synthetic organ-pipe stimuli and three dimensions for a set of wind and string stimuli. Though he did not give a perceptual interpretation of the dimensions, he analyzed the spectral distances[2] between the stimuli with MDSCAL and observed that the spatial solution was similar to that of the dissimilarity ratings.

[2]Differences in energy levels across a bank of 1/3-octave filters.

Wedin and Goude (1972), in another identification experiment, found a three- dimensional model by using factor analysis on dissimilarity ratings of wind and string instruments. Their model revealed a cognitive structure that is in line with the traditional classification into woodwind, brass and string instruments. The physical correlates of the extracted factors were derived from properties of the spectral envelopes: the first factor related to the high strength of upper partials –"sonority" or "overtone richness"; the second factor related to successive intensity-decrease of the upper partials –"dullness" or "overtone poorness"; the third factor related to "low fundamental intensity and an increasing intensity of the first overtones".

3.3 Timbre Spaces

The projection of the stimuli against the MDS axes is called a "timbre space". Miller and Carterette (1975) gave the first example of a timbre space (shown in Figure 3.2) using synthesized tones for studying timbral similarity. They varied the number of harmonics, the amplitude envelope and the onset asynchrony of the harmonics. By using the INDSCAL model they found three dimensions: two of them were related with the number of harmonics; the remaining dimension was related both to the amplitude envelope and onset asynchrony.

Grey (1977) used synthetic sounds based upon an analysis of orchestral instruments. He found that a three-dimensional space (shown in Figure 3.3) was the most useful for interpreting the dissimilarity ratings. The first dimension was associated with the "spectral energy distribution". Though he did not attempt to give a quantitative interpretation, his observations on the nature of distributions were related to measurements of the spectral centroid, spectral spread and spectral skewness. The other two dimensions were related to temporal attributes. The second dimension was associated with the onset synchronicity of the partials during the attack and decay portions of a tone, as well to the overall "spectral fluctuation". The third dimension was related to the noisiness during the attack time –"precedent high frequency", "low amplitude energy" and "inharmonicity energy".

Grey and Gordon (1978) replicated Grey's results and were the first to quantify the dimension related to the spectral energy distribution, by evaluating a set of math-

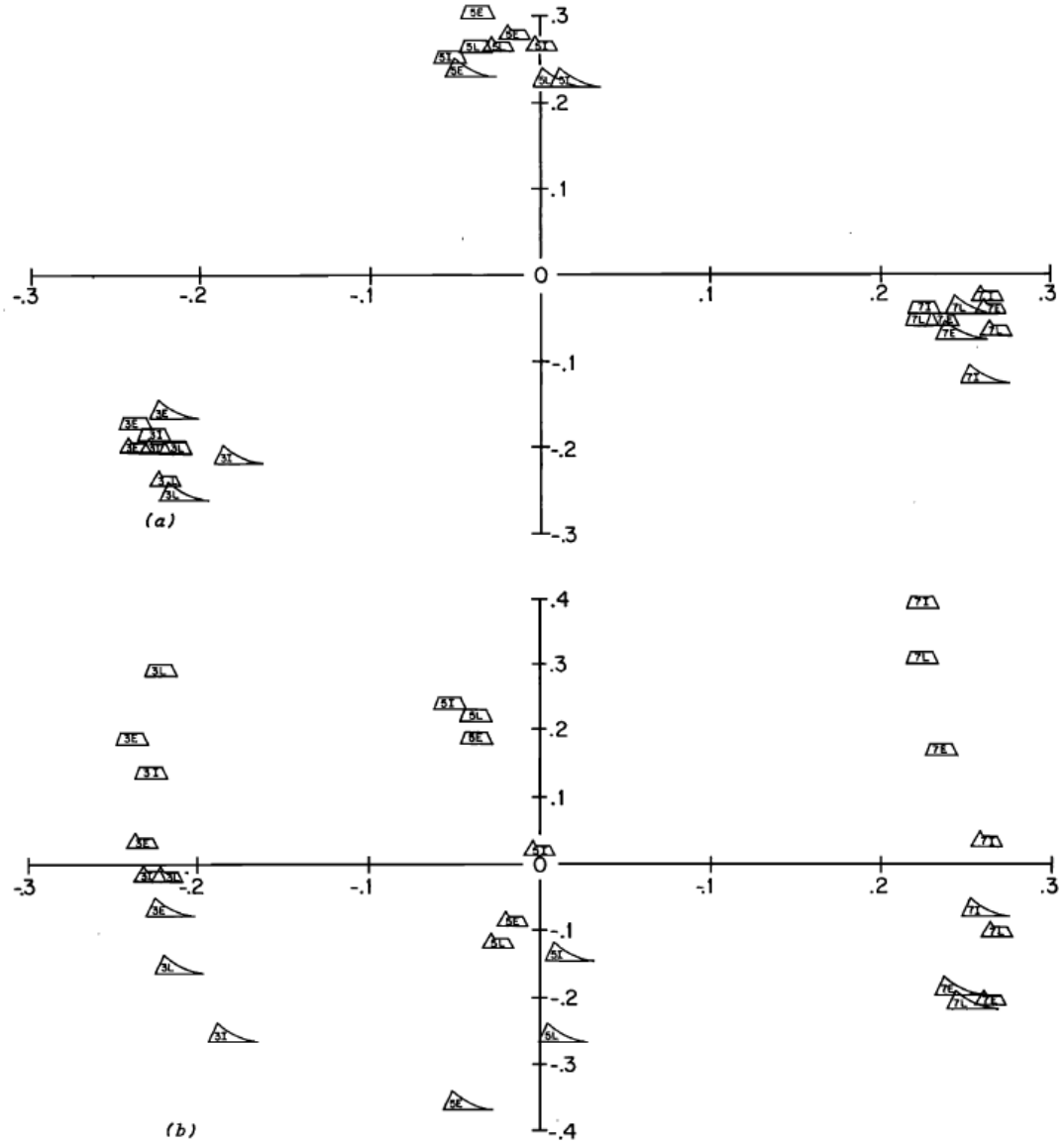


Figure 3.2: A timbre space from Miller and Carterette (1975). Dimension 1 (number of harmonics) on the abscissa is plotted against Dimension 2 (five harmonics versus 3 or 7 harmonics) on the ordinate (a) and against Dimension 3 (envelope) on the ordinate (b). The shape of a point stands for horn, string, or trapezoidal envelope. The pair of letters codes number of harmonics and onset time of harmonics, respectively. Thus, 5E, 5L, 5I stands for a five-harmonic tone with the onset time of the n^{th} harmonic governed by an exponential, a linear and a negative exponential curve respectively. [From Miller and Carterette (1975)]

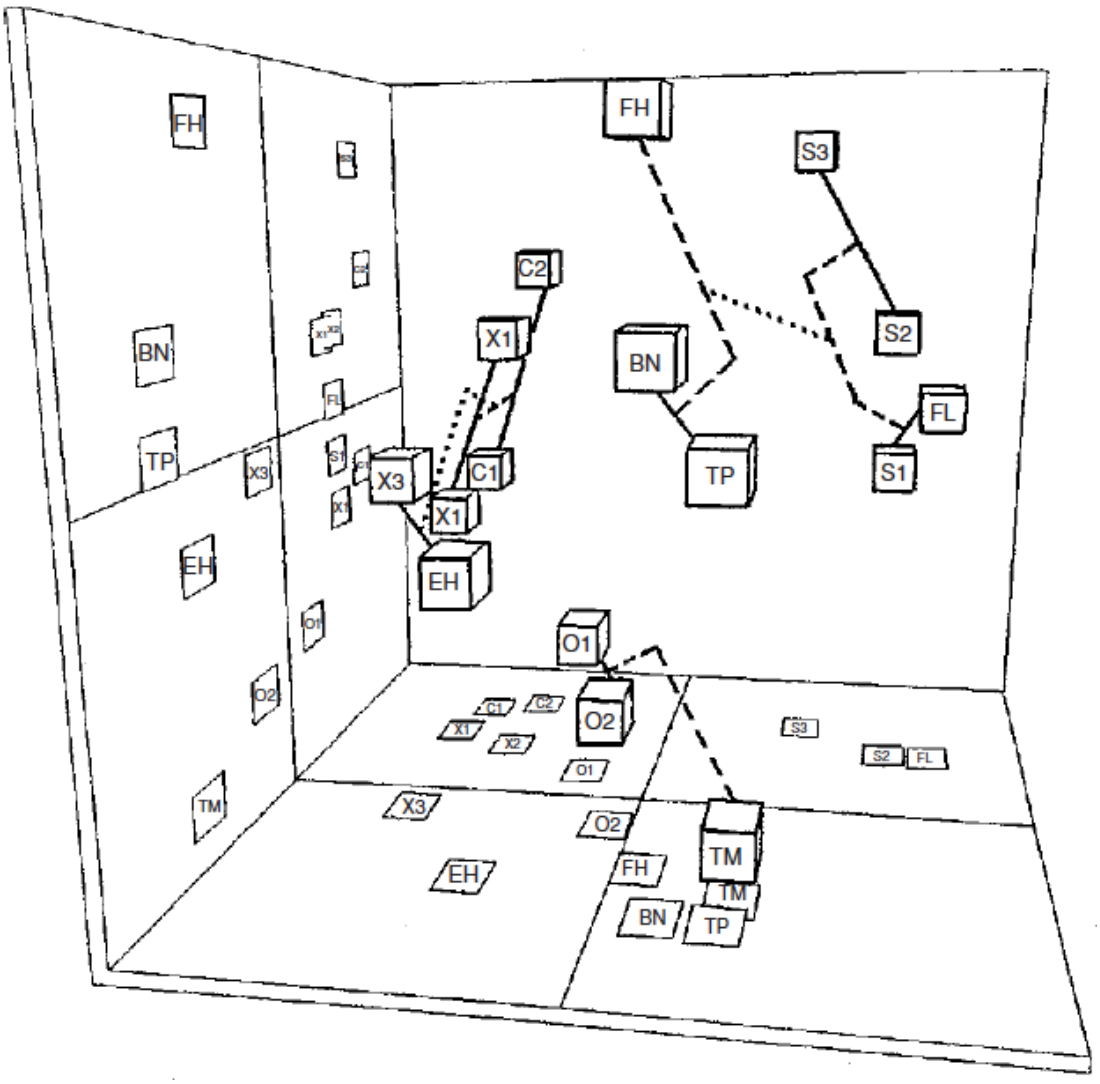


Figure 3.3: Grey's (1977) timbre space. Three-dimensional INDSCAL solution derived from similarity ratings for 16 musical instrument tones. Two-dimensional projections of the configuration appear on the wall and the floor. Abbreviations for the instruments: O1 and O2, two different oboes; C1 and C2, E-flat and bass clarinets; X1 and X2, alto saxophone playing softly and moderately loud, and X3, soprano saxophone, respectively; EH, English horn; FH, French horn; S1, S2, and S3, cello playing with three different bowing styles: *sul tasto, normale, sul ponticello*, respectively; TP, trumpet; TM, muted trombone; FL, flute; BN, bassoon. Dimension 1 (top-bottom) represents spectral envelope or brightness (brighter sounds at the bottom). Dimension 2 (left-right) represents spectral flux (greater flux to the right). Dimension 3 (front-back) represents degree of presence of attack transients (more transients at the front). Hierarchical clustering is represented by connecting lines, decreasing in strength in the order: solid, dashed, and dotted. [From Donnadieu (2007)]

ematical models. The model that correlated most strongly with that dimension was a spectral centroid measure derived from a loudness function.

More systematic attempts to interpret quantitatively perceptual dimensions start with the work of Krimphoff (1993) and Krimphoff, McAdams and Winsberg (1994) based on Krumhansl's (1989) timbre space. Krumhansl used synthetic sounds, created by Wessel, Bristow and Settel (1987), which imitate traditional instruments and hybrids that were synthesized by combing spectrotemporal characteristics of two sounds. Through MDS she found three dimensions, which she related to "Spectral Flux", "Temporal Envelope" and "Spectral Envelope".

Krimphoff et al. (1994) made an acoustic analysis on the sound set used in Krumhansl's study and examined the correlations of various formal models with each axis of that timbre space. The dimensions of "Spectral Envelope" and "Temporal Envelope" correlated strongly (r = 0.94) with spectral centroid and log-attack time respectively. Interestingly, the evaluated spectrotemporal models did not give satisfactory results for interpreting the dimension of "Spectral Flux". Spectral flux, defined by the authors as the RMS variation of the instantaneous spectral centroid over the mean spectral centroid, could only explain 34% of the variance along that dimension. On the contrary, spectral models appeared to be more correlated with that axis: the odd-to-even ratio and spectral deviation accounted for 51% and 72% of the variance respectively.

One of the aims of McAdams' et al. (1995) study was to validate Krimphoff's et al. (1994) quantitative models using a large number of subjects (88) with varying degrees of musical training. They correlated each model with the derived MDS coordinates of 18 sounds drawn from Krumhansl's sound set. The first and second axes of the three-dimensional solution (shown in Figure 3.4) correlated strongly (r=0.94) with log-attack time and spectral centroid as in Krumhansl's study, but spectral deviation did not correlate significantly with the third dimension. Spectral variation[3] gave the highest correlation coefficient for that dimension, but it accounted for only 29% of the variance.

Lakatos (2000) used a broader and a more heterogeneous set of stimuli compared to previous studies, including pitched and unpitched percussive sounds, sustained sounds

[3]Spectral variation is another measurement of spectral flux. It is defined as the average of the correlations between amplitude spectra in adjacent time windows (Krimphoff et al., 1994). See also equation 2.14.

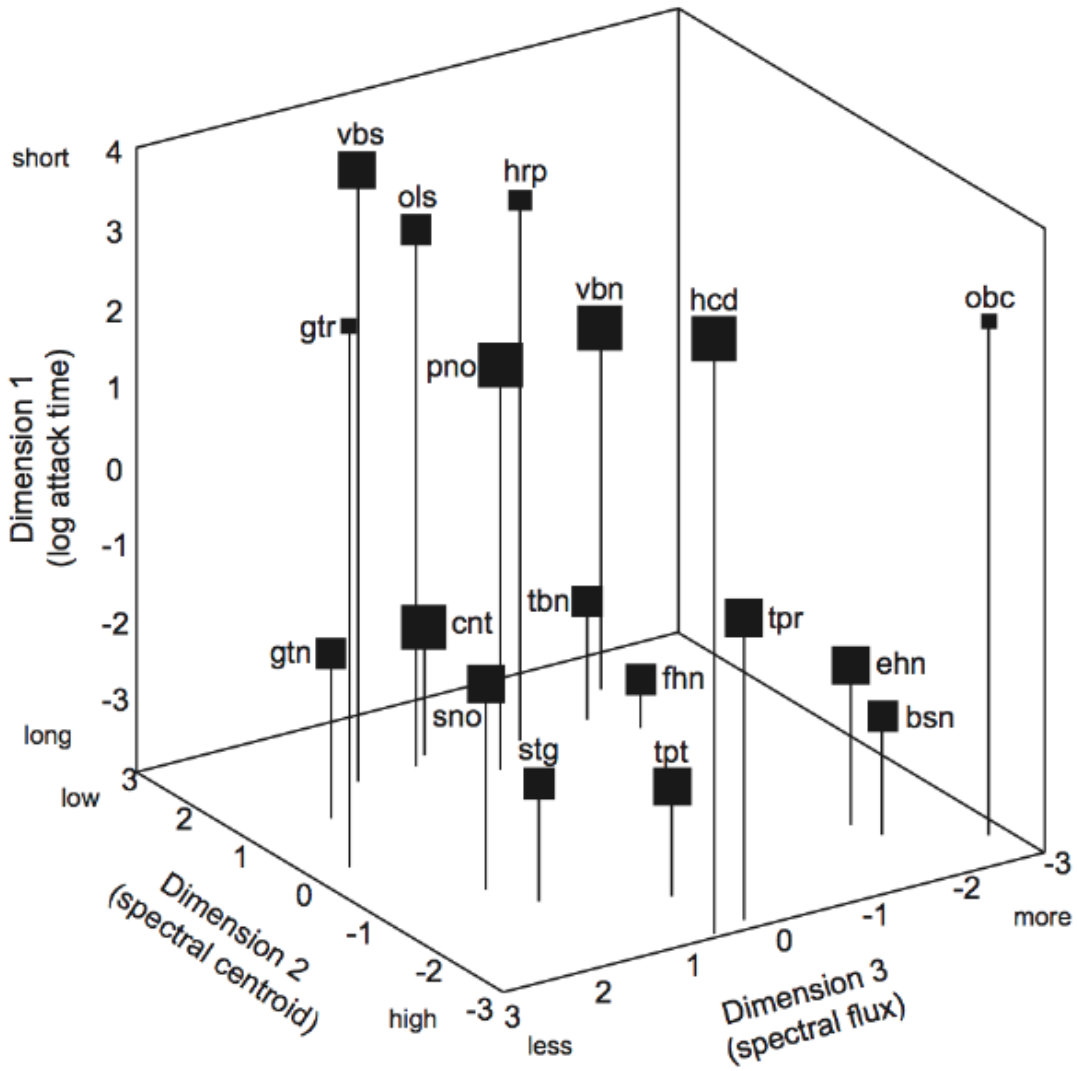


Figure 3.4: McAdam's et al. (1995) timbre space. The CLASCAL solution has three dimensions with specificities (the strength of the specificities is shown by the size of the square). The acoustic correlates of each dimension are also indicated. Abbreviations for the instruments: vbs = vibraphone; hrp = harp; ols = *oboelesta* (oboe\celesta hybrid); hcd = harpsichord; obc = *obochord* (oboe\harpsichord hybrid); gtn = *guitarnet* (guitar\clarinet hybrid); cnt = clarinet; sno = *striano* (bowed string\piano hybrid); ehn = English horn; bsn = bassoon; tpt = trumpet. [From McAdams (2013)]

of pitched orchestral instruments and different modes of excitation. The MDS solution yielded three dimensions for the percussive set and two dimensions for the harmonic and combined set (Figure 3.5). These results confirmed the salience of spectral centroid and log-attack time. The third dimension of the percussive set was associated with "timbral richness" but, as with previous studies, was difficult to interpret psychophysically.

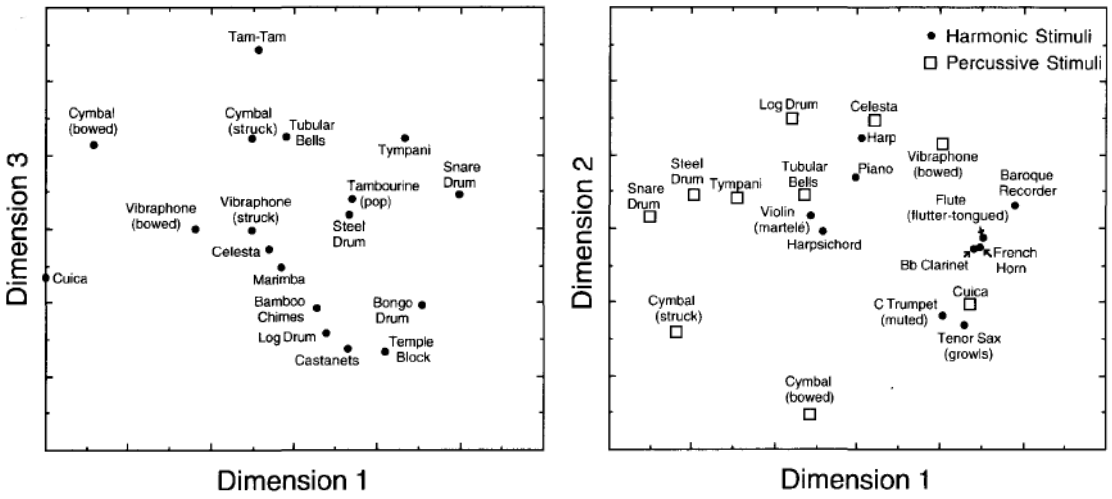


Figure 3.5: Lakatos' (2000) timbre space. CLASCAL solution for the percussive set (a) and the combined set (b). Dimension 1 is correlated with log-attack time, dimension 2 with spectral centroid and dimension 3 with the participants' VAME ratings for timbral "richness". [From Lakatos (2000)]

3.4 Confirmatory Studies

Correlations however, are not proofs of cause-effect relations, so there is a need for confirmatory studies to validate the results of exploratory studies. Grey and Gordon's (1978) results supported the interpretation of the dimension related to spectral shape in Grey's (1977) study. They used half of Grey's stimuli unaltered and made paired modifications on the rest, by exchanging the spectral envelopes between two sounds within each pair while trying to preserve other characteristics. Comparing their results with Grey's timbre space, they observed that the sounds that had exchanged spectral envelopes also exchanged positions along the axis related to "Spectral Energy Distribution". Slight alternations in positions along the other two axes were also observed

since spectral modifications also affected temporal characteristics of the original tones.

The main goal of Caclin, McAdams, Smith, and Winsberg's (2005) study was to validate the interpretation of the "problematic" third dimension against the perceptual saliency of attack time and spectral centroid. They used synthetic tones made up of 20 harmonics with precisely controlled attack time, spectral centroid and spectral irregularity or spectral flux. Spectral irregularity was controlled by attenuating the even harmonics, and spectral flux by a sinusoidal variation of the spectral centroid over the first 100 msec. The dissimilarity judgments confirmed the perceptual saliency of attack time, spectral centroid and spectral irregularity. The effect of spectral flux was tested against the dimensions of attack time and spectral centroid: when all parameters varied concurrently, the effect of spectral flux on the dissimilarity ratings was at best minimal, and was only used to differentiate sounds that had the highest spectral flux values; when both the attack time and spectral centroid were held constant, it was used to differentiate sounds that both had high spectral flux values, or to distinguish sounds that had high versus low spectral flux; when the attack time or spectral centroid were held constant, the effect of spectral flux was more strongly inhibited by attack time than spectral centroid, though it was used by listeners to a much lesser extent than the other two parameters. The authors also noted that their results could be different if spectral flux was present in the sustained portion of the tone, or if it had been modeled differently.

Chapter 4

Verbal Attributes of Timbre

Despite the lack of a specific sound-related vocabulary, we often use language to communicate sound. For example, we can verbally describe an action, or the method of excitation of a sound object (eg. *bowed, struck*), the material of the vibrating object (eg. *metallic*), the temporal and spectrotemporal characteristics of a certain sound (eg. *rustle noise*), and its spectral characteristics (eg. *bright*). Although words often fail to describe the complexity of sounds, their use indicates that essential sound qualities have been recognized.

In this chapter we attempt to create a link between verbal attributes of timbre and a set of audio descriptors. Such an approach can offer insight into acoustical modeling based on semantic and subsequently on perceptual dimensions.

4.1 Previous Studies on Timbre Semantics

Studies on timbre semantics typically use a large number of verbal scales on which subjects rate the stimuli. The goal is usually the elicitation of a number of verbal descriptors, or the identification of semantic dimensions that encompass these descriptors, by using data-reduction techniques such as principal component analysis (PCA) and factor analysis (FA).

According to the semantic differential method (eg. Lichte, 1941; von Bismarck, 1974), the extremes of the scales are labeled by two opposing verbal attributes such as "bright - dull". A potential problem with the semantic differential method is that the bipolar opposites may not be antipodes (Kendall & Carterette, 1993a). A variant of

that method is the verbal attribute magnitude estimation (VAME) according to which the extremes of the scale are labeled by an adjective and its negation such as "bright - not bright" (Kendall & Carterette, 1993a). Other studies (eg. Faure, McAdams & Nosulenko, 1996; Štěpánek, 2006) instead of using a predefined vocabulary, acquire verbal descriptors based on free verbalizations that listeners use to describe timbre differences.

In Lichte's (1941) experiment, subjects judged the dissimilarities of synthetic tones using the scales "rough - smooth", "bright - dull" and "thin - full". In von Bismarck (1974) two groups of subjects comprising musicians and non-musicians rated 35 synthetic sounds on 30 verbal scales, which they had previously chosen themselves from an initial set of 69 scales. A factor analysis on the group of musicians revealed four factors, which were labeled: "dull - sharp", "compact - scattered", "full - empty" and "colorful - colorless". The "dull - sharp" factor was the most prominent accounting for 44% of the variance, while all factors together accounted for 90% of the variance. Pratt and Doak (1976) tested the scales "dull - brilliant", "pure - rich", and "cold - warm" using synthetic sounds. A sine wave was generally described as pure, dull and warm, while sounds with low amplitude on the fundamental frequency were described as rich, brilliant and cold.

In Kendall and Carterette (1993a) subjects rated wind instrument tones on eight bipolar opposites drawn from von Bismarck's experiment: "hard - soft", "sharp - dull", "loud - soft", "complex - simple", "compact - scattered", "pure - mixed", "dim - brilliant" and "heavy - light". Differentiation results using the semantic differential and VAME were poor indicating that the chosen adjectives were inappropriate for rating this type of natural timbres. Kendall and Carterette (1993b) made the same experiment using the VAME method but this time the adjectives were chosen from Piston's (1955) *Orchestration*. By using principal component analysis they found four factors accounting for 90.6% of the variance. Factor 1 was labeled "Power Factor" and was loaded positively by *strong, tense, tremulous, ringing and resonant* and negatively by *smooth, soft, light, weak, and mellow*. Factor 2, labeled "Strident Factor", was loaded positively by *nasal, edgy, and brittle* and negatively by *rich, round, full, warm, and smooth*. Factor 3, the "Plangent Factor" was loaded positively by *ringing, resonant* and negatively by *crisp* and *brilliant*. Factor 4 was labeled "Reed Factor" and was loaded by *reedy, fused, and warm*. They also found a two-dimensional solution using

MDS: the first axis was associated with "nasality - richness" and the second one with "reediness - brilliance".

Faure et al. (1996) elicited a number of verbal descriptors by asking subjects to judge sounds using expressions of the form: "sound 1 is more or less X than sound 2". Dissimilarity ratings were performed on Krumhansl's (1989) sound-set before and after verbalizations. The resulting MDS solutions, before and after verbalizations, were similar, indicating that verbalization did not have an impact on the dissimilarity judgments. Most of the verbal descriptors correlated with more than one dimension, but a few of them correlated only with a single one: *round* was correlated with spectral centroid; *dry* was correlated with log-attack time; *brilliant* and *bright* were correlated with spectral flux. Štěpánek (2006) hypothesized four dimensions of timbre: "gloomy - clear", "harsh - delicate", "full - narrow" and "noisy". The verbal descriptors were collected from spontaneous verbalizations that listeners used to judge the quality of violin and organ-pipe sounds, and from a non-listening experiment that measured the dissimilarity between pairs of verbal attributes. Dislay, Howard and Hunt (2006) used samples of stringed, brass, woodwind and percussive instruments from the MUMS sound library (McGill University Master Samples). Through principal component analysis they found four salient dimensions: "bright, thin, harsh - dull, warm, gentle", "pure, percussive - nasal", "metallic - wooden" and "evolving".

In Zacharakis, Pastiadis and Reiss (2014) English and Greek participants describe musical instrument tones by estimating verbal attribute magnitude values on a predefined set of adjectives presented in their native language. A factor analysis on the two groups of listeners revealed three factors accounting for more than 80% of the variance in the data, which were labeled as: "depth-brilliance" for the Greek group and "brilliance/sharpness" for the English group; "roundness - harshness" for the Greek group and "roughness/harshness" for the English group; "richness/fullness" for the Greek group and "thickness - lightness" for the English group. The inter-correlations of these dimensions between the two groups support the notion of universality of timbre semantics and the different labels were further merged to "luminance", "texture" and "mass". A correlation analysis between semantic dimensions and acoustic parameters associated: "texture" with the energy distribution of partials; *thickness* and *brilliance* with inharmonicity and spectral centroid variation; fundamental frequency with "mass" for the English group, and "luminance" for the Greek group.

4.2 Acoustical Modeling based on Verbal Attributes of Timbre

In the present study we quantify the relationships between verbal attributes of timbre and a set of audio descriptors, having as an ultimate goal to create sounds that exhibit the qualities of a verbal description.

The adjectives, stimuli and listeners' ratings[1] are derived from Zacharakis' et al. (2014) study. The following adjectives are used: *bright, brilliant, cold, compact, dark, deep, dense, dirty, distinct, dry, dull, empty, full, harsh, hollow, light, metallic, mussed, nasal, rich, rough, rounded, sharp, shrill, smooth, soft, thick, thin, warm.*

The sound-set, on which we compute the audio descriptors, consists of 23 sounds with fundamental frequencies in a three-octave range. The following 14 instrument samples are drawn from the MUMS library: *violin, sitar, trumpet, clarinet* and *piano* at A3 (220 Hz); *Les Paul Gibson guitar, baritone saxophone B flat* at A2 (110 Hz); *double bass pizzicato* at A1 (55 Hz); *oboe* at A4 (440 Hz); *Gibson guitar, pipe-organ, marimba, harpsichord* at G3 (196 Hz); *French horn* at A#3 (233 Hz). The rest of the samples are: *flute* at A4; *Acid, Hammond, Moog, Rhodes piano* at A2; *Electric piano (Rhodes), Wurlitzer, Farfisa* at A3; *Bowedpad* at A4.

Figure 4.1 shows the mean ratings of *bright, deep, warm, rounded, dirty* and *metallic*. Listeners performed the ratings on a scale from 0 to 100 and they were free to choose as many adjectives as they felt were necessary for describing most accurately each sound. A mean value of zero on a verbal scale (ex. *sitar's* VAME on *deep*) means that listeners did not choose that scale for describing a specific sound: it indicates that a sound has zero amount of a certain quality, thus zero-values will not be treated as *missing values* in the statistical analysis.

For the audio descriptive analysis we use the median values of a sub-set of harmonic descriptors from Timbre Toolbox, with the number of extracted harmonics set to 20: fundamental frequency, inharmonicity, tristimulus values, odd-to-even ratio, spectral deviation, spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral decrease, spectral roll-off and spectral variation. Spectral slope was discarded in favor of spectral decrease, because it is linearly dependent on the spectral centroid. Harmonic energy and noise energy are parameters that are used to calculate noisiness.

[1] We are using the mean ratings of the English group of listeners.

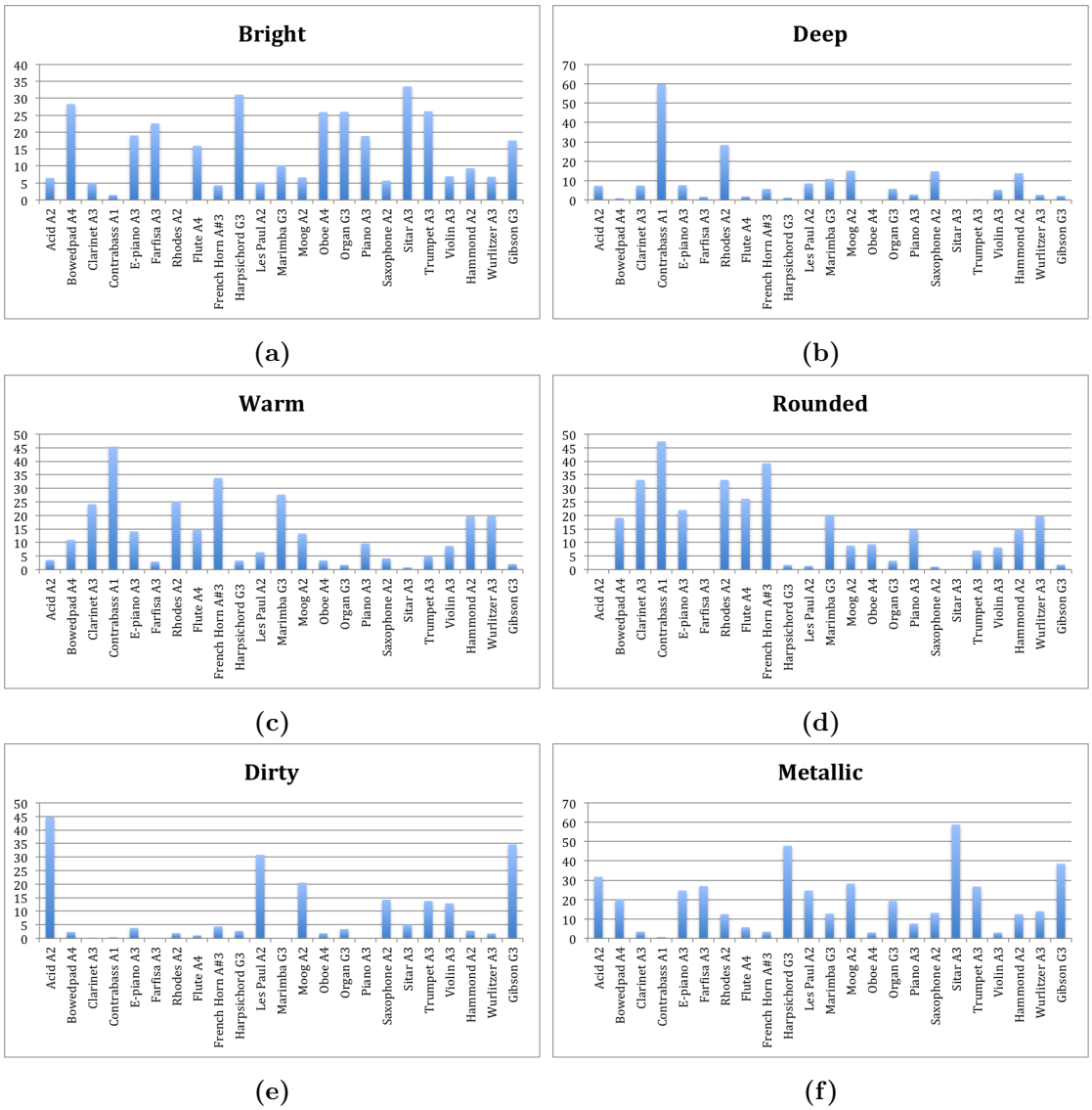


Figure 4.1: A sample of participants' VAME ratings on the scales: (a) Bright; (b) Deep; (c) Warm; (d) Rounded; (e) Dirty; (f) Metallic.

However, the formulation of noisiness in the Timbre Toolbox cannot capture the "true" noise quality of a signal and thus these three descriptors were also discarded.

We chose to use harmonic descriptors, because the sound-set is mainly harmonic, and because they can be directly used as parameters to construct waveforms that have a fixed number of harmonics. By using such waveforms we eliminate the perceptual influence of other sound qualities that are not accounted for by the descriptors used in the analysis. This further enables us to test the following assumption: if audio descriptors do account for certain perceived qualities, then these qualities will also be perceived through simple waveforms, which were constructed according to specific audio descriptor values. All the analyses presented in the next subsections are performed after ranking the data of listeners ratings and the values of audio descriptors.

4.2.1 Correlation Analysis between Verbal Attributes and Harmonic Audio Descriptors

A good starting point for constructing waveforms that exhibit a certain quality would be to inspect the correlations between the adjectives and audio descriptors, shown in Tables 4.1 - 4.3. As an example, a low-pitched sound with strong overtones would be perceived as *rich*: *rich* is negatively correlated with fundamental frequency (r = -0.477) and positively with tristimulus 3 (r = 0.449).

Some adjectives (eg. *full, cold*) are not significantly correlated with any descriptors. This might indicate that listeners used those adjectives inconsistently and spasmodically, or that the audio descriptors used in this analysis cannot capture such qualities.

4.2.2 Predictive Models of Verbal-Attribute Magnitudes

Though correlations are important, relying solely on them might be misleading if there is a high multicollinearity between the independent variables (i.e. the audio descriptors): if there are strong correlations (r>0.8) between the independent variables, as in the present case[2], we cannot decide which of the inter-correlated audio descriptors dominates the perception of the dependent variable (i.e. the verbal description). To

[2]For example, spectral centroid is strongly correlated with spectral spread (r = 0.932) and spectral roll-off (r = 0.957). If we had chosen a different representation, other than the harmonic one, these correlations would be probably weaker.

solve the problem of multicollinearity we are using data reduction techniques, which at the same time can be used to build predictive models based on regression analysis.

Stepwise regression methods select predictors by calculating their statistical contribution in explaining the variance in the dependent variable, and by looking at their semi-partial correlation with the outcome. First, we tested the predictive ability of the *backward elimination* method, the *forward* method and the hybrid *forward - backward elimination* method, using an inclusion criterion of $p<0.05$ and an exclusion criterion of $p>0.1$. As expected, models built with backward elimination explained more of the variance than the other two stepwise methods.

Second, we use predictive models based on principal component analysis (PCA), as this gives rise to mutually orthogonal components that are linear combinations of the predictors. *Partial least squares regression* (Geladi & Kowalski, 1986; Wold, Sjöström & Eriksson, 2001) was preferred over *principal component regression* (PCR) because it maximizes the covariance between predictors and the dependent variable, while PCR may underestimate important predictors because it does not take into account the covariance between them and the dependent variable. The optimal number of components for each dependent variable is selected according to the *robust component selection* statistic (RCS), which combines the goodness-of-fit and the predictive ability of the model (Engelen & Hubert, 2005).

Table 4.4 shows the variances explained using the backward elimination versus the partial least squares regression (PLSR). As can be seen, PLSR performs better in most of the cases. Figures 4.2 and 4.3 demonstrate the predicted magnitude of each sound on the scales of *bright, deep, warm, rounded, dirty* and *metallic* according to PLSR, against the participants' ranked ratings. The beta regression coefficients for every scale are reported in tables 4.5 - 4.7.

4.2.3 Conclusions

Preliminary results based on judging and predicting the qualities of synthetic tones show that clusters of audio descriptors, as well as their relative values within the cluster, account for sound qualities expressed by the adjectives. The tones were constructed by altering the properties of sawtooth-waveforms according to a family of audio descriptors, which was indicated by inspecting the correlations between them and the verbal attributes. For example, to construct a palette of *clear* and *not clear* sounds, we used

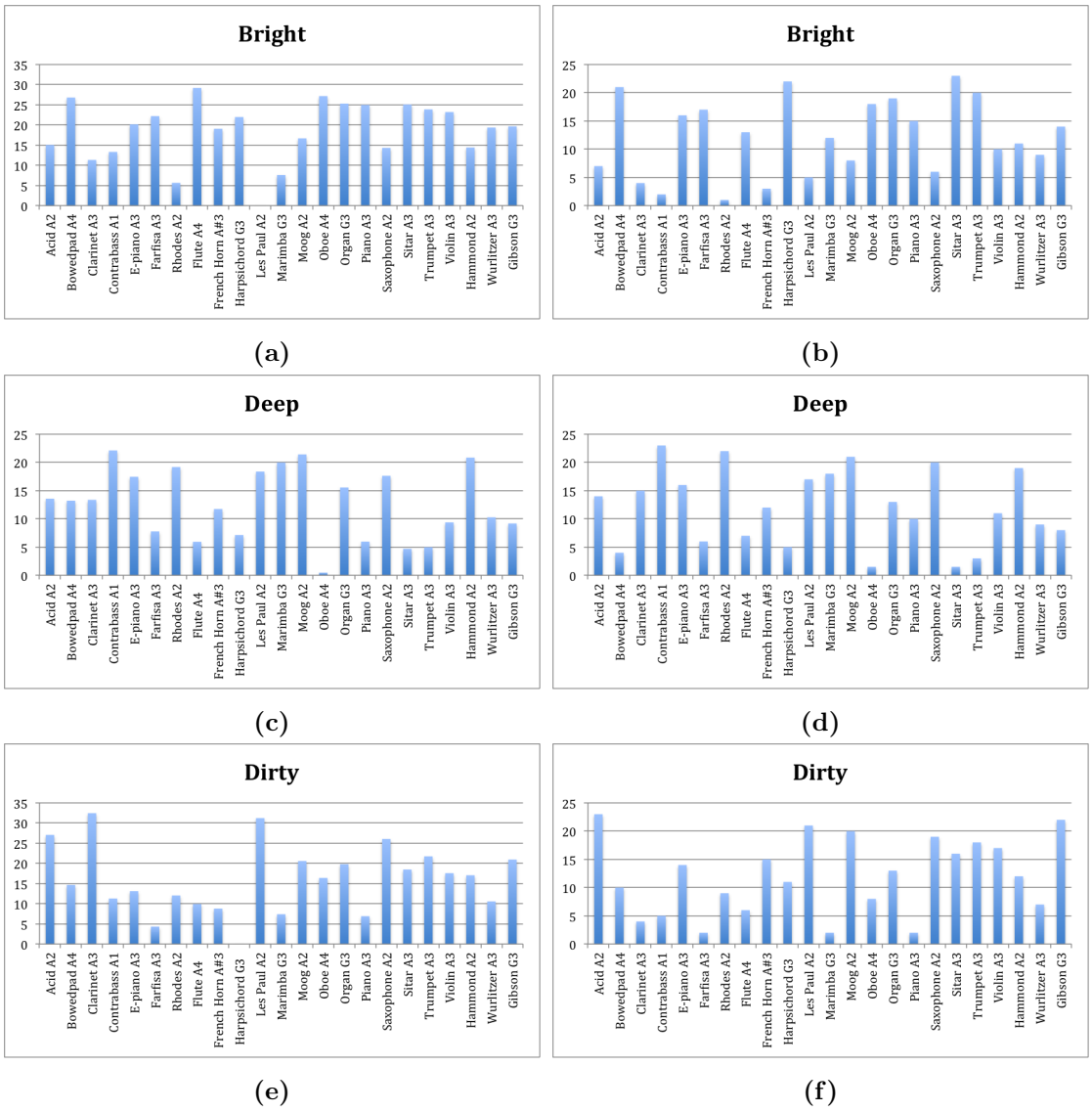


Figure 4.2: (a), (c), (e): predicted verbal magnitude based on PLSR. (b), (d), (f): participants' ranked ratings.

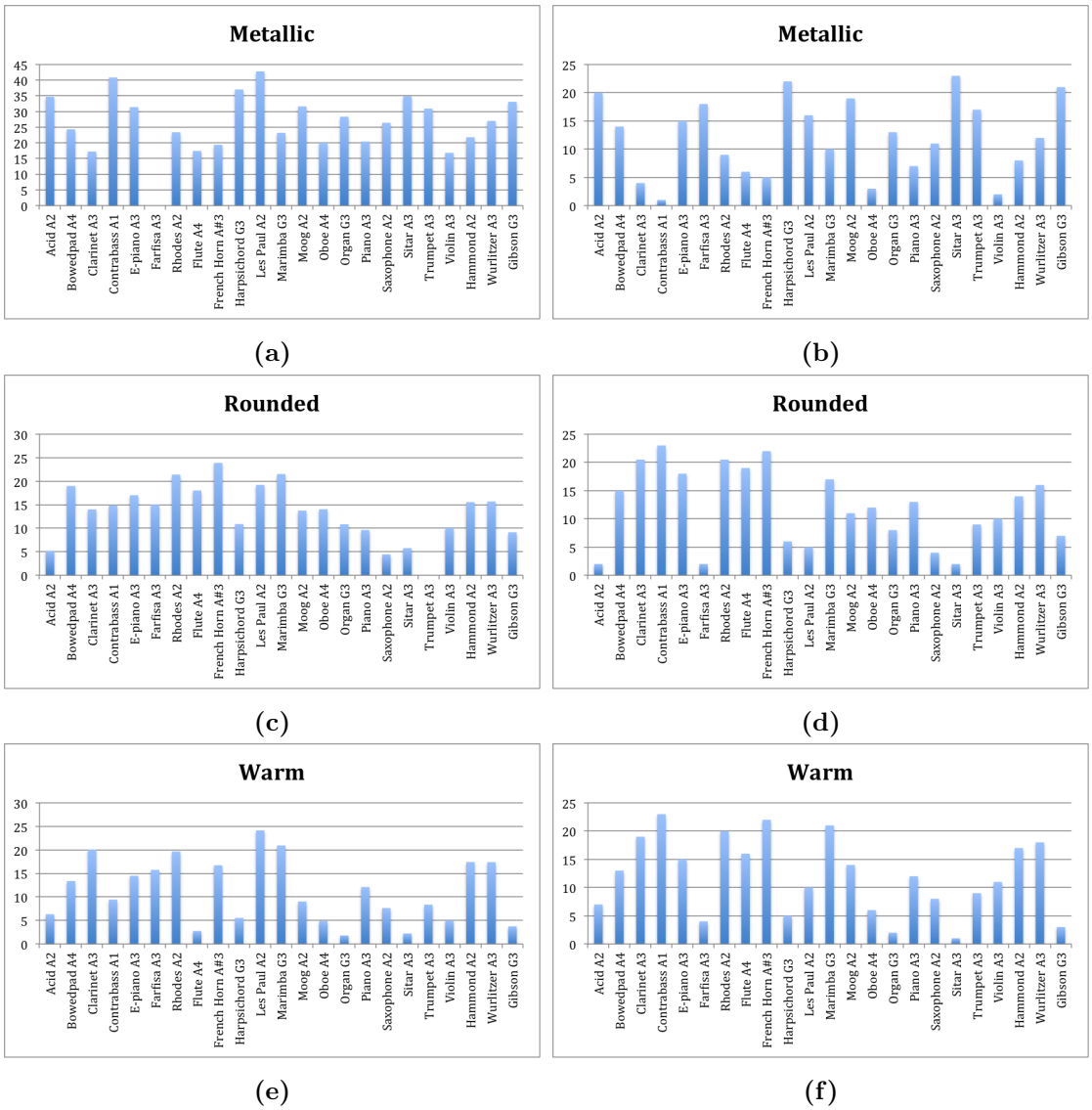


Figure 4.3: (a), (c), (e): predicted verbal magnitude based on PLSR. (b), (d), (f): participants' ranked ratings.

sawtooth-waveforms with different fundamental frequencies and different amounts of inharmonicity[3].

Depending on the adjectives, audio descriptors in general form different multidimensional spaces in which sounds are located according to their verbal-attribute magnitudes. Furthermore, based on our analysis/resynthesis scheme, PLSR proves to be a usefull tool for predicting the magnitude that each verbal attribute has on a given sound. Therefore, it is possible to create, classify and order sounds using audio parameters that correspond to verbal descriptions.

[3]Inharmonicity was added by varying the inharmonicity coefficient according to the formula: $f_n = nf_o\sqrt{(1+an^2)}$, where n is the harmonic and a the inharmonicity coefficient.

Audio Descriptors	Bright	Brilliant	Clear	Cold	Compact
Fundamental Frequency	.479*		.567**		
Inharmonicity	-.525*		-.501*		
Tristimulus 1					
Tristimulus 2					
Tristimulus 3					
Deviation					-.497*
Odd/Even		-.449*			
Centroid	.465*	.492*			
Spread	.463*	.483*			
Skewness					
Kurtosis					
Decrease					
Roll Off		.438*			
Variation					
	Dark	Deep	Dense	Dirty	Distinct
Fundamental Frequency	-.794**	-.756**	-.578**		
Inharmonicity	.658**	.649**			
Tristimulus 1				-.504*	
Tristimulus 2					
Tristimulus 3					
Deviation					
Odd/Even	.419*	.446*			
Centroid	-.480*	-.651**			
Spread	-.473*	-.606**			
Skewness					
Kurtosis					
Decrease				.494*	
Roll Off		-.574**			
Variation				.487*	

Table 4.1: Correlations. *p<0.05, **p<0.01.

Audio Descriptors	Dry	Dull	Empty	Full	Harsh
Fundamental Frequency					
Inharmonicity		.514*			
Tristimulus 1			.527**		-.602**
Tristimulus 2					
Tristimulus 3			-.486*		.625**
Deviation					
Odd/Even					
Centroid		-.523*			.496*
Spread		-.507*			.482*
Skewness					-.536**
Kurtosis					-.544**
Decrease			-.503*		.448*
Roll Off		-.493*			.570**
Variation					
	Hollow	Light	Metallic	Mussed	Nasal
Fundamental Frequency		.555**			
Inharmonicity		-.510*			
Tristimulus 1	.449*	.495*		-.499*	-.644**
Tristimulus 2					
Tristimulus 3	-.540**				.708**
Deviation					.436*
Odd/Even					
Centroid					.647**
Spread					.589**
Skewness	.482*				-.567**
Kurtosis	.509*				-.539**
Decrease		-.514*		.512*	.542**
Roll Off					.708**
Variation				.450*	

Table 4.2: Correlations. *p<0.05, **p<0.01.

Audio Descriptors	Rich	Rough	Rounded	Sharp	Shrill
Fundamental Frequency	-.477*				
Inharmonicity					-.525*
Tristimulus 1		-.664**	.545**	-.433*	-.466*
Tristimulus 2					
Tristimulus 3	.449*	.567**	-.485*		.441*
Deviation					
Odd/Even					-.508*
Centroid				.436*	.714**
Spread				.418*	.715**
Skewness			.461*		-.538**
Kurtosis			.477*		-.522*
Decrease		.640**	-.454*		
Roll Off		.418*		.421*	.700**
Variation					
	Smooth	Soft	Thick	Thin	Warm
Fundamental Frequency			-.571**		
Inharmonicity			.417*		
Tristimulus 1	.514*	.554**			.586**
Tristimulus 2					
Tristimulus 3		-.532**			-.451*
Deviation					
Odd/Even					.511*
Centroid				.516*	-.559**
Spread				.565**	-.535**
Skewness		.508*			.476*
Kurtosis		.495*			.469*
Decrease	-.436*	-.458*			-.489*
Roll Off				.542**	-.585**
Variation					

Table 4.3: Correlations. *p<0.05, **p<0.01.

	Variance Explained (%)				
Method	Bright	Brilliant	Clear	Cold	Compact
BCKWD	60.6	56.8	63.1	0	85.8
PLSR	79.8	78.2	72.7	0.95	74.2
	Dark	Deep	Dense	Dirty	Distinct
BCKWD	83.2	87	91.5	83.3	35.3
PLSR	83.5	90.7	82.5	85.5	54.8
	Dry	Dull	Empty	Full	Harsh
BCKWD	70.5	46.9	46.9	61.1	71
PLSR	50.2	76	49.4	60.5	70.7
	Hollow	Light	Metallic	Mussed	Nasal
BCKWD	71.2	73.5	0	69	69.4
PLSR	78.5	87.5	95.4	77.9	61.3
	Rich	Rough	Rounded	Sharp	Shrill
BCKWD	63.8	74.1	53.1	52.6	51.2
PLSR	60.6	48.4	89.7	81.3	49.5
	Smooth	Soft	Thick	Thin	Warm
BCKWD	58.3	85.8	58.8	61.7	61.9
PLSR	61.6	95.9	56.1	80.9	99

Table 4.4: Variance explained by backward elimination (BCKWD) and partial least squares regression (PLSR).

Audio Descriptors	Regression Coefficients (β)				
	Bright	Brilliant	Clear	Cold	Compact
Fundamental Frequency	.47	.55	.53	0	-1.03
Inharmonicity	-.11	.05	.01	0	.16
Tristimulus 1	.17	-.43	.14	-.05	-.29
Tristimulus 2	.07	.01	-.15	.06	-.26
Tristimulus 3	-.38	-1.01	-.44	-.02	-.15
Deviation	-.33	-.48	-.23	-.13	-.68
Odd/Even	-.56	.07	-.60	-.06	-.05
Centroid	.15	.03	.12	.01	.31
Spread	.29	.21	.03	-.03	.25
Skewness	-.11	-.51	-.15	.03	.30
Kurtosis	-.19	-.64	-.1	.02	.45
Decrease	-.14	.37	-.17	.05	.1
Roll Off	.02	-.20	-.18	-.01	.45
Variation	.37	.30	-.24	.19	-.71
	Dark	Deep	Dense	Dirty	Distinct
Fundamental Frequency	-.32	-.25	-.41	.12	0
Inharmonicity	.24	.03	-.06	.28	-.23
Tristimulus 1	.03	-.01	.07	-0.36	-.07
Tristimulus 2	-.16	-.39	-.15	.21	-.08
Tristimulus 3	.15	-.01	-.01	.47	.05
Deviation	.07	.58	.34	.37	-.85
Odd/Even	.14	.09	-.15	1.28	-.25
Centroid	-.17	-.21	-.11	.48	0
Spread	-.11	-.30	0	-.25	-.03
Skewness	-.12	.07	-.1	.17	-.30
Kurtosis	-.15	.06	-.13	.2	-.34
Decrease	-.04	.08	.01	.34	-.03
Roll Off	-.01	-.27	-.06	.27	-.07
Variation	.02	.28	.56	.55	-.31

Table 4.5: Beta coefficients of partial least squares regression.

Audio Descriptors	Regression Coefficients (β)				
	Dry	Dull	Empty	Full	Harsh
Fundamental Frequency	.07	.01	.08	-.23	-.01
Inharmonicity	-.05	.03	-.06	-.18	-.04
Tristimulus 1	-.05	.06	.27	.25	-.18
Tristimulus 2	.03	.06	-.01	-.12	-.05
Tristimulus 3	.26	-.10	-.16	-.05	.14
Deviation	-.62	-.06	.29	.52	-.31
Odd/Even	-.18	.02	-.10	-.14	-.06
Centroid	.07	-.08	-.10	-.18	.16
Spread	.08	-.09	-.01	-.14	.12
Skewness	-.16	.11	-.01	-.21	-.07
Kurtosis	-.08	.10	0	-.21	-.08
Decrease	-.04	-.04	-.25	-.20	.09
Roll Off	.09	-.09	-.13	-.22	.16
Variation	-.42	-.04	-.15	.10	.37
	Hollow	Light	Metallic	Mussed	Nasal
Fundamental Frequency	.33	.30	-.63	-.20	-.05
Inharmonicity	-.04	-.23	-.88	.20	-.01
Tristimulus 1	.07	.31	-.50	-.29	-.23
Tristimulus 2	-.13	.42	.09	.45	.05
Tristimulus 3	-.11	-.21	-.49	.06	.34
Deviation	.52	-.35	-.84	-.20	-.11
Odd/Even	.27	-.01	.69	.31	.33
Centroid	.09	.14	-.92	-.02	.16
Spread	-.02	.13	-.10	.15	.12
Skewness	.26	-.01	.75	.28	.02
Kurtosis	.30	-.04	-.77	.23	.06
Decrease	.01	-.43	.63	.47	.18
Roll Off	.01	.10	.85	.27	.29
Variation	-.03	.48	.48	.34	-.10

Table 4.6: Beta coefficients of partial least squares regression.

Audio Descriptors	Regression Coefficients (β)				
	Rich	Rough	Rounded	Sharp	Shrill
Fundamental Frequency	-.21	-.07	0	.10	.14
Inharmonicity	-.28	.01	-.20	-.41	-.14
Tristimulus 1	-.28	-.14	.17	-.08	-.10
Tristimulus 2	-.52	-.04	-.01	0	.01
Tristimulus 3	.16	.12	-.22	-.11	0
Deviation	.24	-.09	.73	-.31	.10
Odd/Even	-.13	-.03	-.05	-.31	-.10
Centroid	-.10	.07	-.19	.03	.15
Spread	-.45	.04	-.20	.08	.10
Skewness	-.07	-.05	.11	-.11	.01
Kurtosis	-.09	-.06	.11	-.09	.02
Decrease	.33	.11	-.02	.03	.11
Roll Off	-.36	.09	-.25	.02	.11
Variation	.15	.25	-.11	-.41	.15
	Smooth	Soft	Thick	Thin	Warm
Fundamental Frequency	.04	-.30	-.29	.28	.69
Inharmonicity	-.22	-.33	.21	.04	.33
Tristimulus 1	.23	.10	-.02	-.10	-.36
Tristimulus 2	.21	.03	-.17	.07	-.11
Tristimulus 3	-.18	-.74	.17	.05	-.39
Deviation	.35	.20	.08	-.56	.40
Odd/Even	-.12	-.53	.11	.16	.73
Centroid	-.13	-.06	-.11	.28	.11
Spread	-.13	-.27	-.07	.28	-.62
Skewness	.03	.10	-.14	.05	-.66
Kurtosis	.04	-.11	-.16	.07	-.39
Decrease	-.15	-.10	0	-.03	-.03
Roll Off	-.17	-.31	-.06	.32	-.70
Variation	.02	.04	.08	.13	-.36

Table 4.7: Beta coefficients of partial least squares regression.

Chapter 5

Audio Descriptive Synthesis

The conclusions drawn from past experiments and the analysis made in the previous chapter revealed how audio descriptors interact with each other, and how they relate and account for perceived sound qualities. That was an important first step that we had to take before start making effective use of the audio descriptors in a synthesis context. As Jean-Claude Risset points out:

> "So, in order to profit from the immense sound resources offered by the computer, it becomes necessary to develop a psychoacoustical science, involving a knowledge of the correlations between the physical parameters and the perceptible characteristics of sound." (Risset, 1971)

Audio descriptive synthesis (AUDESSY) makes use of audio descriptors along additive synthesis. Additive synthesis is a malleable technique for constructing sounds based on a set of partials, which are precisely defined in terms of their frequency and amplitude envelopes, and onset (or offset) synchrony (or asynchrony). Furthermore, the wide range of operations that can be applied to sets of partials makes additive synthesis attractive to composers. Most common operations include: time stretching or compression; changing the spectral density by adding or removing partials; pitch transposition by preserving the frequency spacing between partials; expansion or compression in spectral space by altering the frequency spacing between the partials; spectral tuning by adjusting the partials' frequencies to match a predetermined spectrum.

Additive synthesis is accomplished using SPEAR. First, we specify in a matrix the number of partials, their time-varying amplitude and frequency values, and total

duration. These values are then exported in the proper text format (shown in Figure 2.4) and imported to SPEAR. SPEAR will synthesize the final sound using a bank of oscillators that interpolate linearly (in frequency and amplitude) between every time frame.

Audio descriptors would normally measure the effect that such operations have on the resulting spectrum, but in AUDESSY such cause-effect relations are either eliminated or reversed: audio descriptors are used as global spectrum modulators (or shapers) and pose structural constraints that allow us to control the higher-level organization of the partials. Thus, AUDESSY could be summarized in the following steps:

1. Specification of the source-spectrum in terms of partials and their time varying amplitude and frequency values.

2. Specification of the target-morphology in terms of audio descriptors.

3. Optional modulation of a single or multiple audio descriptors using as a carrier the source-spectrum.

4. Synthesis of the final sound according to steps two and three while retaining as much as possible the properties of the source.

In chapter 4 we used AUDESSY to create sounds according to a verbal description. Previous approaches to sound synthesis based on verbal descriptions (Ethington & Punch, 1994; Gounaropoulos & Johnson, 2006) have not investigated systematically the relationships between adjectives and perceived sound qualities. Most importantly, the qualities that were recognized and attributed to partials' relations were not quantified.

Other attempts made to provide synthesis-control over timbral features are more related to navigation between sounds in a feature space and thus they tend to emphasize on the construction of hybrid tones rather than new ones (Hourdin, Charbonneau & Moussa, 1997; Haken, Fitz & Christensen, 2007; Jehan & Schoner, 2001; Le Groux, 2006). Some other approaches focus on the resynthesis of sounds according to a very limited set of audio descriptors and other sonic parameters, but they do not take into account their inter-dependencies (Jensen, 1999; Park, Biguenet, Li, Richardson & Scharr, 2007; Hoffman & Cook, 2007). For instance, altering the spectral centroid of a sound will also affect its spectral spread, unless certain constraints are used.

AUDESSY gains control over the stability and variability of the synthesis parameters using optimization. In other words, it uses constraints that allow some parameters to vary while others are held as much invariant as possible.

5.1 Optimization

Optimization is a useful tool when one needs to make a single "best" decision through a plethora of available choices. Xenakis (1992) used optimization based on *linear programming* to compose the pieces *Duel* and *Stratégie*. In the paragraph related to the analysis of *Duel* he wrote:

> "It appeals to relatively simple concepts: sonic constructions put into mutual correspondence by the will of the conductors, who are themselves conditioned by the composer." (Xenakis, 1992, p. 113)

More recently, optimization is used in computer-aided orchestration where the goal is usually to find the best instrumental combination that approximates a target sound (Rose & Hetrick, 2009; Carpentier & Bresson, 2010;). An optimization scheme is a necessity in AUDESSY since we are dealing with an underdetermined system: there are fewer constraints (i.e. audio descriptors) than unknowns (i.e. partials' amplitudes) and the system has an infinite number of solutions. For instance, there is infinite number of combinations of partials amplitudes for a given spectral centroid.

More specifically, we use the *sequential quadratic programming* (SQP) method implemented in MATLAB, to solve the following problem: find the amplitude values p_h that minimize the sum of partials' amplitudes for each time frame, using as constraints the audio descriptors. For instance, if we use the spectral centroid as a constrain, the problem will be formulated as follows:

Minimize: $\sum_{h=1}^{H} p_h$, where p_h is the amplitude of partial h and H is the total number of partials.

Subject to: $SC = \sum_{h=1}^{H} f_h p_h$, where f_h is the frequency of partial h and SC the target value of the spectral centroid.

With the additional constrain $0 \leq p_h \leq 1$ and an initial vector $\mathbf{P_0}$.

The SQP method will find *local* rather than *global* solutions because it relies heavily on the supplied initial vector $\mathbf{P_0}$, which is in our case the initial amplitude values of the partials (step 1 in the previous paragraph). Thus, this allows us to come up with the "best" sound: if there is a feasible solution and the initial spectrum satisfies all constraints we will get back the same spectrum unaltered while if not, we get a spectrum that is as much similar as possible to the initial one while having all of the constraints satisfied.

5.2 Plausible Uses of AUDESSY

In this example, we present how AUDESSY can be used to construct a sound from scratch. First, we specify the duration, fundamental frequency and calculate the number of harmonic partials using as an upper limit the Nyquist frequency. The amplitude of each partial is calculated by the function that defines a sawtooth wave. Shimmer and inharmonicity are added for each partial by using a tendency mask with a uniform probability distribution, a lower bound of '0' and an upper bound of '0.5' that falls linearly to zero. Thus, we start from a noisy signal that gradually becomes inharmonic and finally a perfect sawtooth wave. We apply to the final spectrum the amplitude envelope of a piano sound (Figure 5.1), and we use SPEAR to synthesize the result (Figure 5.2). Figure 5.3 shows the effect of the above operations on spectral flux (shown as the variation of spectral centroid). We shape further the spectrum by applying a lower and constant spectral centroid at 700 Hz. Figure 5.4 shows the result of the optimization: the structure of the partials in the frequency space is maintained, with the lower ones being significantly strengthened and the upper ones being attenuated.

Another plausible use of AUDESSY is related with timbre spaces. Timbre spaces can be used to achieve *timbral transpositions* based on *timbral intervals*. A timbral interval can be considered as a vector having a specific magnitude and orientation that connects two different timbres inside a timbre space (Figure 5.5).

Ehresman and Wessel (1978) were the first to test if listeners can perceive timbral analogies in a two-dimensional timbre space. They found that the interval between timbres A and B would be perceived as analogous to another interval between timbres

C and D if the vectors $\overrightarrow{AB}$ and $\overrightarrow{CD}$ have a similar magnitude and orientation. McAdams and Cunibile (1992) tested further the vector model in the three-dimensional space from Krumhansl (1989) by comparing timbral transpositions based on vectors that had: right magnitude and right direction with respect to a reference vector; right magnitude and wrong direction; wrong magnitude and right direction; wrong magnitude and wrong direction.

Though the main result globally supported the predictive ability of the model, the specificities that were present in the stimulus set distorted the transposed interval vectors and therefore the subjective impression of timbral analogies. Therefore, timbral transpositions may be more applicable to homogeneous timbre spaces, constituting of synthesized sounds or blended combinations of several acoustic instruments (McAdams, 2013).

AUDESSY can be used to create uniformly spaced sounds by controlling the effect of every perceptual dimension. Furthermore, on a given sound-set, the ideal sound for achieving an accurate timbral transposition usually does not exists. With AUDESSY we can fill the space by creating sounds that match the ideal timbre space coordinates for a given timbral interval, and encapsulate, as much as possible, the properties of the nearest sound-neighbors to the target points.

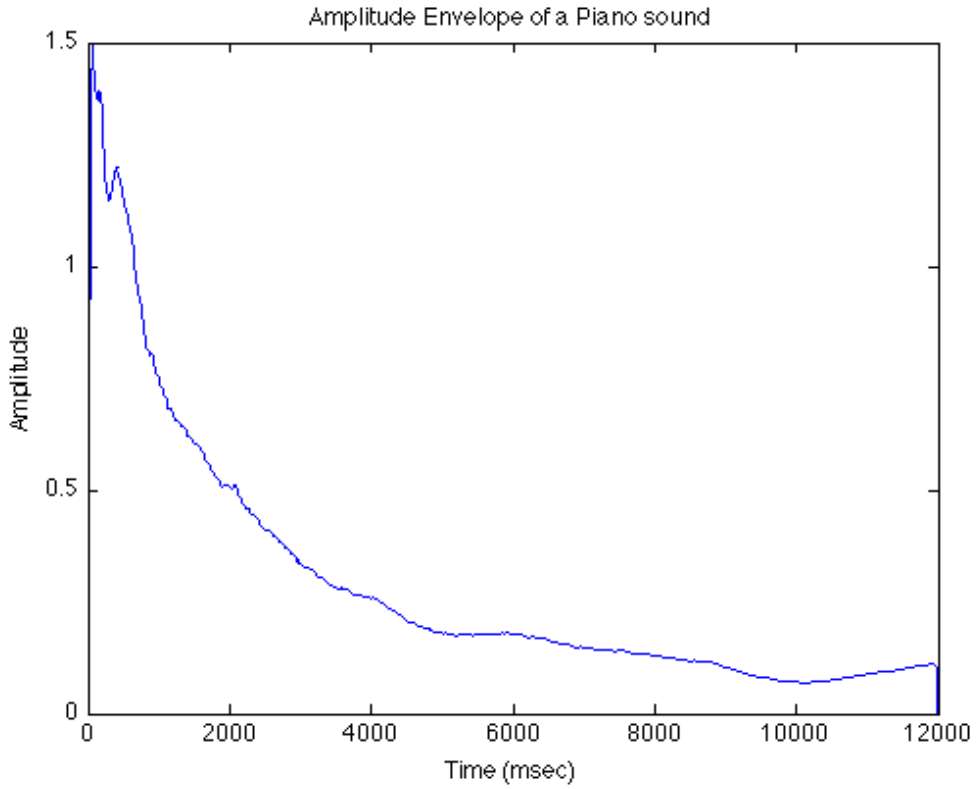


Figure 5.1: Amplitude envelope of a piano sound.

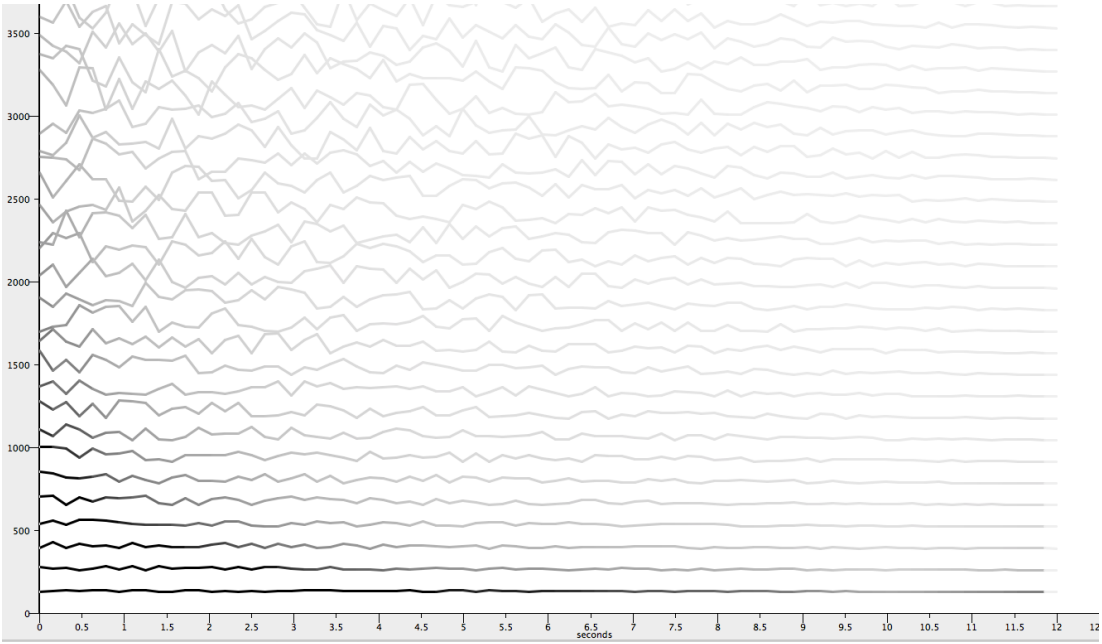


Figure 5.2: Synthesis without controlling the spectral centroid.

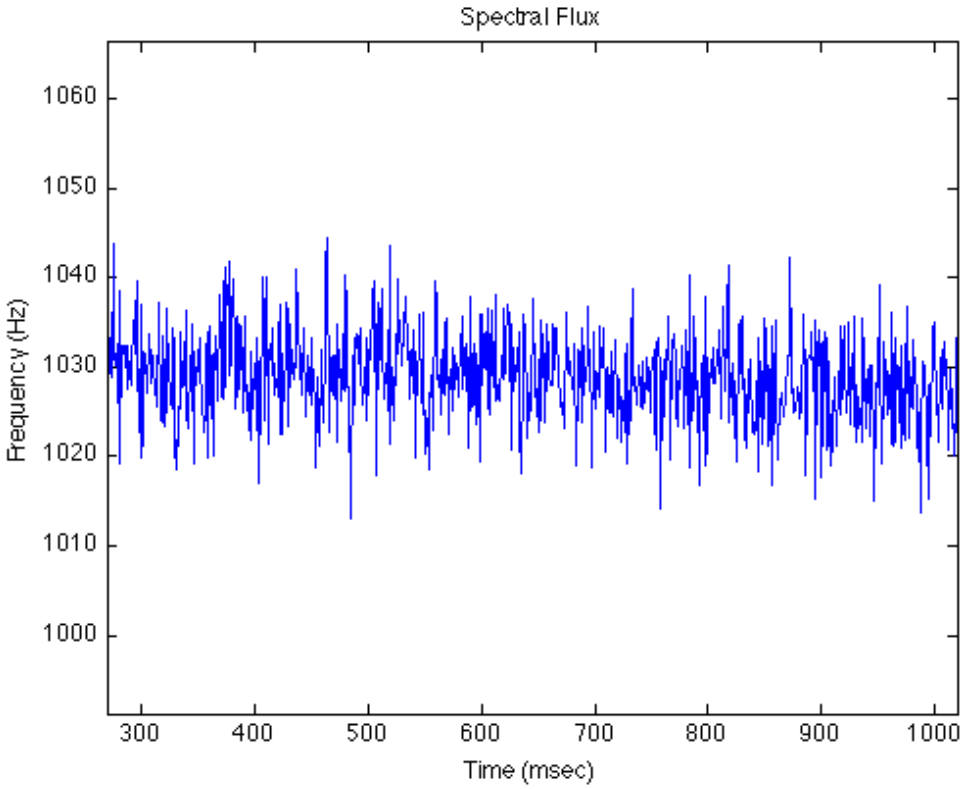


Figure 5.3: Spectral flux as a result of the above operations.

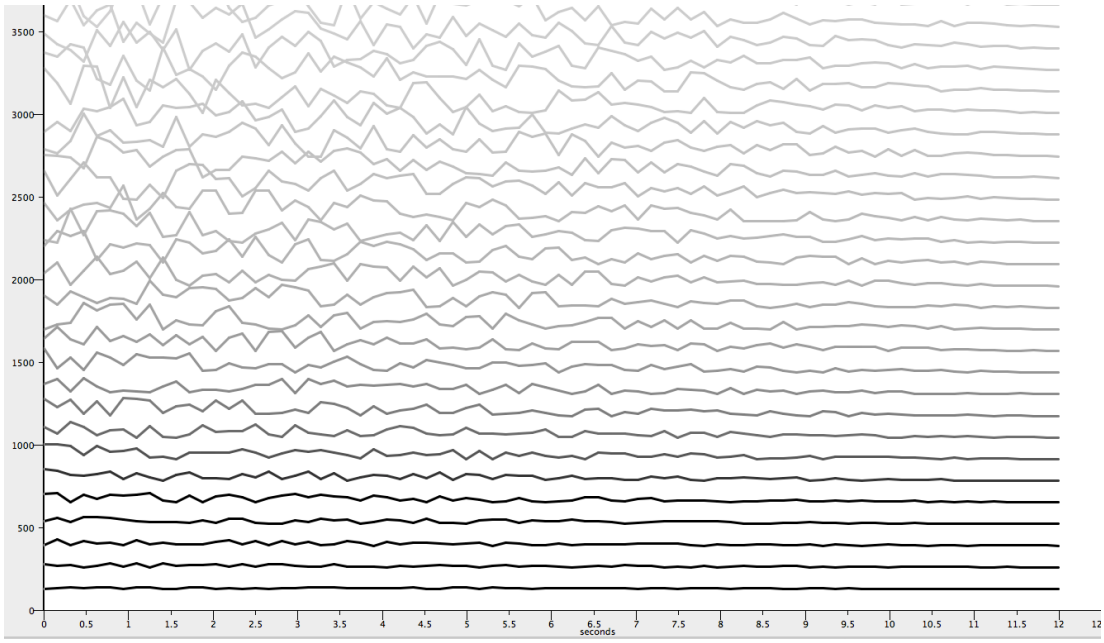


Figure 5.4: Synthesis with a fixed spectral centroid at 700 Hz.

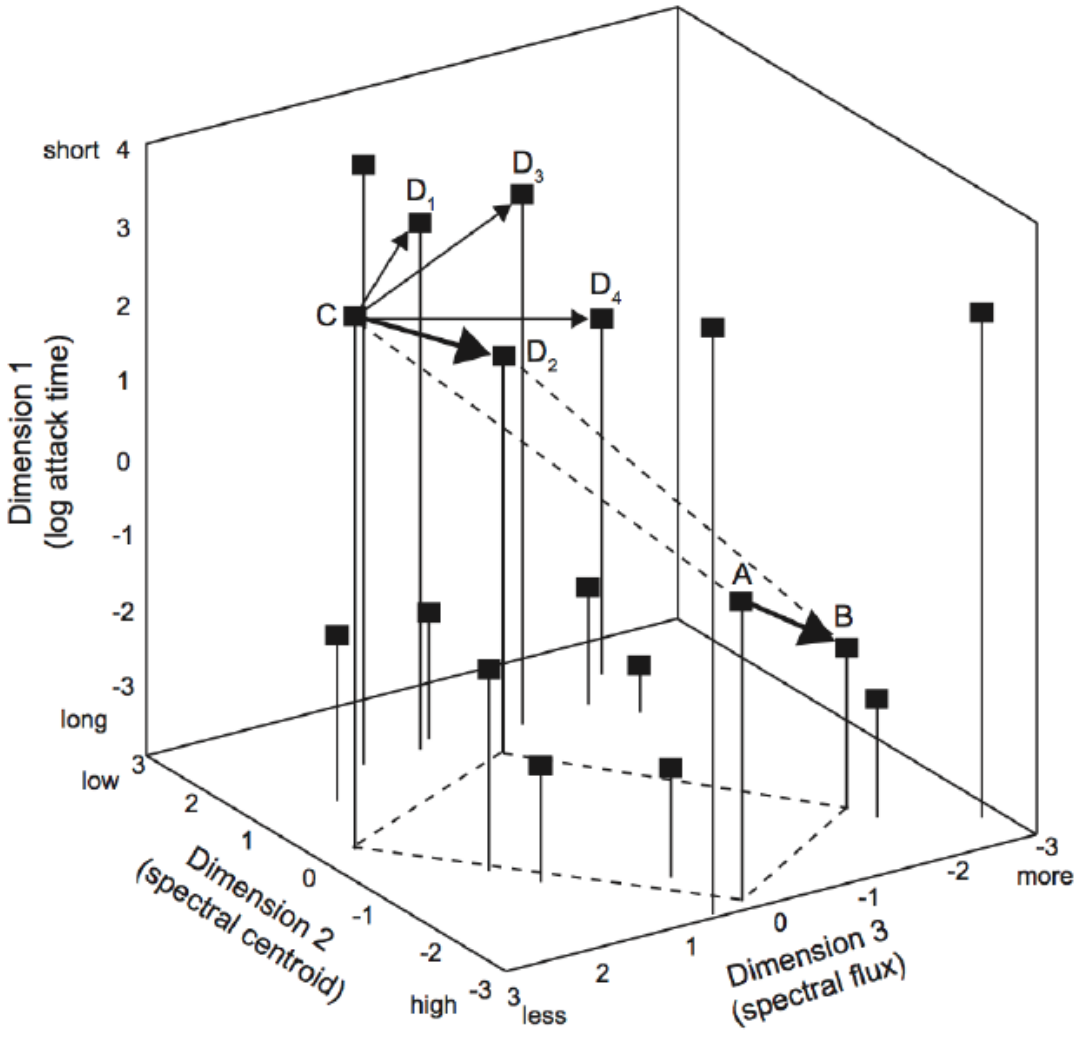


Figure 5.5: Examples of timbral intervals in a timbre space. The aim is to find an interval starting with C and ending on a timbre D that resembles the interval between timbres A and B. If we present timbres D_1-D_4 the vector model would predict that listeners would prefer D_2, because the vector CD_2 is the closest in length and orientation to that of AB. [From McAdams (2013)]

References

ASA. (1960). *American standard acoustical terminology.* New York: American Standards Association.

Berg, P. (2012). *Lecture notes.* Institute of Sonology, The Hague.

Berger, K. W. (1964). Some factors in the recognition of timbre. *Journal of the Acoustical Society of America*, *36*(10), 1888-1891.

Bregman, A. (2001). *Auditory scene analysis.* 2nd ed. MIT Press.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, *118*, 471-482.

Carpentier, G., & Bresson, J. (2010). Interacting with symbolic, sound and feature spaces in orchidée, a computer-aided orchestration environment. *Computer Music Journal*, *34*(1), 10-27.

Disley, A. C., Howard, D. M., & Hunt, A. D. (1996). Timbral description of musical instruments. *Proceedings of the 9th International Conference on Music Perception and Cognition (ICMPC 09)*, (pp. 61 -68).

Donnadieu, S. (2007). Mental representation of the timbre of complex sounds. In J. W. Beauchamp (Ed.), *Analysis, synthesis, and perception of musical sounds.* NewYork: Springer.

Ehresman, D., & Wessel, D. L. (1978). Perception of timbral analogies. *Rapports de l'IRCAM*, (Vol.13). Paris, France: IRCAM-Centre Pompidou.

Engelen, S., & Hubert, M. (2005). Fast model selection for robust calibration methods. *Analytica Chimica Acta*, *544*, 219-228.

Erickson, R. (1975). *Sound structure in music.* Berkeley, CA: University of California Press.

Faure, A., McAdams, S., & Nosulenko, V. (1996). Verbal correlates of perceptual dimensions of timbre. *Proceedings of the 4th International Conference on Music Perception and Cognition (ICMPC4)*, (pp. 108-116). McGill University, Montreal,

Canada.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, *185*(1), 1-17.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, *61*(5), 1270-1277.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, *63*(5), 1493-1500.

Haken, L., Fritz, K., & Christensen, P. (2007). Beyond traditional sampling synthesis: Real-time timbre morphing. In J. W. Beauchamp (Ed.), *Analysis, synthesis, and perception of musical sounds.* NewYork: Springer.

Hourdin, C., Charbonneau, G., & Moussa, T. (1997). A sound-synthesis technique based on multidimensional scaling of spectra. *Computer Music Journal*, *21*(2), 56-68.

Jehan, T., & Schoner, B. (2001). An audio driven perceptually meaningful timbre synthesizer. *Proceedings of the 2001 International Computer Music Conference*.

Jensen, K. (1999). *Timbre models of musical sounds.* Unpublished doctoral dissertation, University of Copenhagen.

Kendall, R. A., & Carterette, E. C. (1993a). Verbal attributes of simultaneous wind instrument timbres: 1. von Bismarck's adjectives. *Music Perception*, *10*(4), 445-468.

Kendall, R. A., & Carterette, E. C. (1993b). Verbal attributes of simultaneous wind instrument timbres: 2. adjectives induced from Piston's 'Orchestration'. *Music Perception*, *10*(4), 469-502.

Klingbeil, M. K. (2009). *Spear: Spectral analysis, editing, and resynthesis: Methods and applications.* Unpublished doctoral dissertation, Columbia University.

Krimphoff, J. (1993). *Analyse acoustique et perception du timbre.* unpublished DEA thesis, Université du Maine, Le Mans, France.

Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caractérisation du timbre des sons complexes. 2 : Analyses acoustiques et quantification psychophysique. [characterization of the timbre of complex sounds. 2: Acoustic analysis and psychophysical quantification]. *Journal de Physique*, *4*(C5), 625-628.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In S. Nieltzén & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music.* Amsterdam: Excerpta Medica.

Kruskal, J. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*, 115-129.

Lagrange, M., Badeau, R., David, B., Bertin, N., Echeveste, J., Derrien, O., et al. (2010). The desam toolbox: Spectral analysis of musical audio. *International Conference on Digital Audio Effects, Graz*.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, *62*(7), 1426-1439.

Lartillot, O., & Toiviainen, P. (2007). A matlab toolbox for musical feature extraction from audio. *International Conference on Digital Audio Effects, Bordeaux*.

Le Groux, S. (2006). *Mapping high-level sonic percepts to sound generation.* Unpublished doctoral dissertation, Universitat Pompeu Fabra, Barcelona, Spain.

Lichte, W. H. (1941). Attributes of complex tones. *Journal of Experimental Psychology*, *28*(6), 455-480.

McAdams, S., & Cunibile, J.-C. (1992). Perception of timbral analogies. *Philosophical transactions of the royal society, London, Series B*, *336*, 383-389.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*, 177-192.

Miller, J. R., & Carterette, E. C. (1975). Perceptual space for musical structures. *Journal of the Acoustical Society of America*, *58*, 711-720.

Park, T. H., Biguenet, J., Li, Z., Richards, C., & Scharr, T. (2007). Feature modulation synthesis (fms). *Proceedings of the 2007 International Computer Music Conference*.

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).

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio features from musical signals. *Journal of the Acoustical Society of America*, *130*, 2902-2916.

Peeters, G., McAdams, S., & Herrera, P. (2000). Instrument description in the context of mpeg-7. *Proceedings of the 2000 International Computer Music Conference, Berlin*, (pp. 166-169). San Francisco, CA: International Computer Music Association.

Piston, W. (1955). *Orchestration.* New York: W. W. Norton & Co.

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. F. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing.* Sijthoff, Leiden.

Plomp, R. (1976). Timbre of complex tones. In R. Plomp (Ed.), *Aspects of tone*

sensation: A psychophysical study. London, Academic Press.

Pratt, R., & Doak, P. (1976). A subjective rating scale for timbre. *Journal of Sound and Vibration*, *45*, 317-328.

Randall, J. K. (1967). Three lectures to scientists. *Perspectives of New Music*, *5*(2), 124-140.

Risset, J.-C. (1971). Synthesis of sound by computer and problems concerning timbre. *La Revue Musicale*(268-269), 117-128.

Risset, J.-C. (1991). Timbre analysis by synthesis: Representation, imitation and variants for musical composition. In G. de Poli, A. Piccialli, & C. Roads (Eds.), *Representations of musical signals.* London: The MIT Press.

Rose, F., & Hetrick, J. E. (2009). Enhancing orchestration technique via spectrally based linear algebra methods. *Computer Music Journal*, *33*(1), 32-41.

Smalley, D. (1994). Defining timbre - refining timbre. *Contemporary Music Review*, *10*, 35-48.

von Bismarck, G. (1974). Timbre of steady tones: A factorial investigation of its verbal attributes. *Acustica*, *30*, 146-159.

Štěpánek, J. (2006). Musical sound timbre: Verbal descriptions and dimensions. *International Conference on Digital Audio Effects, Montreal*.

Wedin, L., & Goude, G. (1972). Dimension analysis of the perception of instrumental timbre. *Scandinavian J. Psychol*, *13*, 228-240.

Wessel, D., Bristow, D., & Settel, Z. (1987). Control of phrasing and articulation in synthesis. *Proceedings of the 1987 International Computer Music Conference, Champaign/Urbana*, (pp. 108-116). San Francisco, CA: International Computer Music Association.

Winsberg, S., & Carroll, D. (1989). A quasi-nonmetric method for multidimensional scaling via an extended euclidean model. *Psychometrika*, *54*, 217-229.

Wold, S., Sjostrom, M., & Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(1), 109-130.

Xenakis, I. (1992). *Formalized music.* (Rev. ed). New York: Pendragon Press.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception: An Interdisciplinary Journal*, *31*(4), 339-358.