1.4 A Pure Data Spectro-Morphological Analysis Toolkit for Sound-Based Composition

Gilberto Bernardes Sound and Music Computing Group, INESC TEC, Portugal

Matthew E. P. Davies Sound and Music Computing Group, INESC TEC, Portugal

Carlos Guedes Sound and Music Computing Group, INESC TEC, Portugal; NYU Abu Dhabi, Uniteted Arab Emirates

Abstract

This paper presents a computational toolkit for the real-time and offline analysis of audio signals in Pure Data. Specifically, the toolkit encompasses tools to identify sound objects from an audio stream and to describe sound objects attributes adapted to music analysis and composition. The novelty of our approach in comparison to existing audio description schemes relies on the adoption of a reduced number of descriptors, selected based on perceptual criteria for sound description by Schaeffer, Smalley and Thoresen. Furthermore, our toolkit addresses the lack of accessible and universal labels in computational sound description tasks by unpacking applied terminology from statistical analysis and signal processing techniques. As a result, we improve the usability of these tools for people with a traditional music education background and expand the possibilities for music composition and analysis.

Keywords: Sound morphology, Content-based audio processing, Audio descriptors, Concatenative sound synthesis.

1. Introduction

Sound description is an essential task in many disciplines from phonetics and psychoacoustics to musicology and audio processing, which address it for a variety of purposes and through very distinct approaches (Gouyon et al., 2008). Among these, the computational approach to sound description is the most relevant to our work.

Computational approaches to sound description have gained increased attention in recent years given the large expansion of multimedia content over personal and public databases and the consequent need for effective algorithms for browsing, mining, and retrieving these huge collections of multimedia data (Grachten et al., 2009). Music creation has rapidly followed these approaches and has been incorporating these tools into composition processes (Humphrey, Turnbull, & Collins, 2013).

The output quality of content-based audio processing systems is commonly dependent on the audio representations they adopt, because most processing relies on such data. The most common approach to represent audio in such systems is the adoption of *audio descriptors*, which measure properties of audio signal content. For example, the brightness of an audio sample can be extracted by the audio descriptor spectral centroid, which measures the “center of mass” of the spectral representation of an audio signal.

Most content-based audio processing systems that extensively use audio descriptors tend to prevent users from accessing them. For example, the use of audio descriptors in applications like Shazam [1] and Moodagent [2] take place during the implementation phase of the algorithm and are hidden from the system’s interface. However, creative applications like Echo Nest Remix API [3], CataRT (Schwarz, 2006) and earGram (Bernardes et al., 2013) give users access to audio descriptors and even encourage them to experiment with their organization in order to retrieve and generate different audio sequences and outputs.

However, even if many computationally-extracted audio descriptors—in particular those computed from audio data by simple means, commonly referred to as low-level audio descriptors—measure musical or perceptual properties of sound, they are not adapted to the terminology of musical practice and are meaningless without understanding the underlying statistical analysis and signal processing techniques. Therefore, one can conclude that one of the most evident and prominent barriers for operating audio descriptors in creative music applications is the lack of accessible and meaningful labels adapted to particular application contexts and user preferences. By unpacking the terminology, we believe that the usability of content-based audio systems can increase considerably, and appeal to a larger audience, most-importantly including musicians.

Inspired by the work of Ricard (2004), Peeters and Deruty (2008), and Schnell, Cifuentes, and Lambert (2010), our goal is to develop a computational toolkit for real-time and offline segmentation and description of sound objects [4] in Pure Data (Puckette, 1996) targeted toward users more familiar with music theory and practice than with music technology. Furthermore, contrary to the recent tendency in music information retrieval (MIR) to adopt large numbers of audio features in content-based audio processing systems, our toolkit purposefully encompasses a very limited number of descriptors and aligns with some recent studies by Mitrovic, Zeppelzauer, and Eidenberger (2007) and Peeters et al. (2011), which have shown that the information expressed by the totality of audio descriptors developed exposes a high degree of redundancy. To this end, we will rely on criteria of musical perception grounded in sound-based theories by Schaeffer (1966), Smalley (1986, 1997), and Thoresen (2007) to select an appropriate set of computational audio descriptors for music analysis and composition.

The remainder of this paper is organized as follows. Section 2 introduces our research as well as the grounding principles of three musicological sound-based theories, which we summarize along with their criteria for sound description in Section 3. Section 4 presents the computational strategies implemented for identifying sound objects automatically. Section 5 details at length the proposed sound descriptors included in our toolkit. Section

6 briefly presents musical applications that adopt our toolkit. The paper concludes with conclusion in Section 7.

2. A Musicological Approach to Sound Description

Until the 1940s, music composition was only confined to acoustic instrumental and vocal models and highly tied to the concept of the musical note. Within this paradigm, pitch and rhythm were understood as the primary musical elements of musical structure with timbre (restricted almost exclusively to orchestration) and other attributes of sound thought of as secondary (Thoresen, 2007).

The appearance of new electronic instruments and sound manipulation strategies by that time broke the paradigm linking sound to the physical object producing it, and allowed composers to work with dimensions that were previously inaccessible or totally disregarded in music composition, particularly the use of all sonic phenomena as raw material for composition. In electroacoustic music, the basic structural unit of the composition is no longer the musical note. Instead, the concept of the sound object comes to the fore, significantly extending the spectrum of possibilities (from note to noise) and the exploration of timbre as a compositional strategy.

As a result, much electroacoustic music was particularly resistant to traditional analysis and categorization. In addition, the new dimensions explored in electroacoustic music existed for some decades without any theoretical ground or formal definition that could articulate the relevant shift within musical composition. Clearly, a unique set of terms and concepts was needed to discuss, analyze, and interpret electroacoustic music (Smalley, 1986).

In the early years of electroacoustic music theory, the discourse was largely monopolized by engineering terminology, consequently lacking theoretical and aesthetic reflection. Schaeffer's *Traité des Objets Musicaux* (TOM) was the first substantial treatise on the subject, which addressed the correlation between the world of acoustics and engineering with that of the listener and musical practice. While the technology used by Schaeffer is now old, his overall approach taken to listening, description, characterization and organization of sound is still a reference.

For more than four decades, Schaeffer's TOM had no deep implications in (electronic) musical analysis (Thoresen, 2007; Landy, 2007). This neglect is commonly attributed to the difficulty of Schaeffer's writings and their unavailability in English until 2004. In this regard, a note should be paid to the work of Chion (1983), who systematized Schaeffer's work, as well as the spectromorphology and aural sonology theories by Smalley (1986, 1997) and Thoresen (2007) that acknowledge and reformulate Schaeffer's morphological criteria within a simpler yet more concise and operable framework. In what follows, we briefly detail the guiding principles of Schaeffer's morphological criteria of sound perception and then delve into their definition. Whenever pertinent, we will interleave the definition of Schaeffer's morphologic criteria with perspectives from Smalley and Thoresen towards the ultimate goal of establishing a theoretical basis to support our toolkit.

3. Schaeffer's *Solfeggio* and After

In TOM, Schaeffer reframes the act of listening to sound by articulating a phenomenological theory that is primarily concerned with the abstracted characteristics of sounds, rather than their sources and causes (Chion, 1983); an attitude that he refers to as reduced listening. This listening attitude ultimately establishes the basis of a *solfeggio* for sound-based works, or in Schaeffer's words, a "descriptive inventory which precedes musical activity" (Schaeffer, 1966, as cited in Chion, 1983, p. 124).

Schaeffer's *solfeggio* is divided into five stages, of which the first two, commonly addressed together as typo-morphology, are those most relevant to our work. These two stages aim (i) to identify sound objects from an audio stream, then (ii) to classify them into distinctive types, and, finally, (iii) to describe their morphology. Typology takes care of the first two operations and morphology the third.

Despite the lack a systematic musicological approach for identifying sound objects in the sound continuum (Smalley, 1986), it is important to understand the conceptual basis of these unitary elements, as it is a required processing stage prior to their morphological description. A sound object can be identified by its particular and intrinsic perceptual qualities that unify it as a sound event on its own and distinguishes it from all other sound events (Chion, 1983).

The morphological criteria are defined as "observable characteristics in the sound object" (Chion, 1983, p. 158), and "distinctive features [...or] properties of the perceived sound object" (Schaeffer, 1966, p. 501), like the mass of a sound (e.g. sinusoidal or white noise), sound's granularity and dynamics.

Two main concepts, matter and form, organize Schaeffer's morphology. For Schaeffer, if matter refers to the characterization of stationary spectral distributions of sound, then sound matter is what we would hear if we could freeze the sound in time. Form exposes the temporal evolution of the matter.

Matter encompasses three criteria: mass, harmonic timbre, and grain. Mass is the "mode of occupation of the pitch-field by the sound" (Schaeffer, 1966, as cited in Chion, 1983, p. 159). By examining the spectral distribution of a sound object, it is possible to define its mass according to classes that range from noise to a pure sinusoidal sound.

Harmonic timbre is the most ambiguous criterion presented in Schaeffer's morphology. Its definition is very vague and closely related to the criterion of mass—complementing it with additional qualities of the mass (Schaeffer, 1966). Smalley (1986) avoids this criterion altogether and Thoresen (2007) presents a sound descriptor, spectral brightness, that clearly belongs to the harmonic timbre criterion within the mass criteria (sound spectrum according to Thoresen's terminology).

Grain defines the microstructure of the sound matter, such as the rubbing of a bow. Even though it describes an intrinsic temporal dimension of the sound, it is under the criterion of matter because it examines a micro time scale of music, which the human hear does not distinguish as separate entities (Schaeffer, 1966).

Sound shape/form encompass two criteria: dynamic, and pace. The dynamic criterion exposes and characterizes the shape of the amplitude envelope. Schaeffer distinguished several types of dynamic profile (e.g. unvarying, impulsive, etc.), as well as several types of attack (e.g. smooth, steep, etc.).

The pace (*allure* in French) is another ambiguous concept presented in Schaeffer's TOM defined as fluctuations in the sustain of the spectrum of sound objects—a kind of generalized vibrato. Smalley avoids this criterion. Thoresen (2007) adopts the criterion and enlightens its definition by providing simpler, yet reliable categories for describing both the nature (pitch, dynamic, and spectral), and the quality of possible undulations (e.g. their velocity and amplitude). Still, we find Thoresen's definition of pace unsystematic, in light of a possible algorithmic implementation, since it does not offer a concise description of the limits of the criteria.

4. Identifying Sound Objects Computationally

We adopt two main methods to identify and segment an audio stream into sound objects: onset detection and beat tracking algorithms.

Onset detection algorithms aim to find the location of notes or similar sound events in the audio continuum by inspecting the audio signal for sudden changes in energy, spectral energy distribution, pitch, etc. Current algorithms for onset detection adopt quite distinct audio features, or combinations of them, in order to convey improved results for specific types of sounds, such as percussive, pitched instrumental, or soundscapes (Bello et al., 2004). Figure 1 illustrates the adoption of two functions given a monophonic pitched audio input on which we obtain different results when inspecting for onsets. While the drops in the amplitude function group events into larger segments, the different "steps" in the continuous pitch tracking function provide a better means for the task of segmenting the given example into notes events.

In order to address a variety of sounds, we adopted three distinct onset detection methods in our framework. The first is a perceptually based onset detection algorithm by Brent (2011), which intends to be used for pitched (polyphonic) sounds. The second algorithm is based on a pitch-tracking method by Puckette, Appel and Zicarelli (1998) and aims to detect note onsets of monophonic pitched audio. The third is an adaptive onset detection function largely based on the work of Brossier (2006), which inspects the audio for spectral changes. A particular feature of this method, which intends primarily to be used for environmental sounds, is its ability to adapt to the local properties of the signal to improve the onset estimates.

Beat tracking is a computational task that aims to automatically find an underlying tempo and detect the locations of beats in audio files. It corresponds to the human action of tapping a foot on perceptual music cues that reflect a locally constant inter-beat interval. Although beats cannot always be considered as sound objects, according to Schaeffer's theory, we adopt these temporal units because of their relevancy under music with a strong beat, such as electronic dance music.

A comprehensive description of our audio beat tracking algorithm—largely based on Dixon (2007)—is out of the scope of this paper. For a comprehensive description of all segmentation strategies applied in the toolkit please refer to Bernardes (2014, p. 45-49).

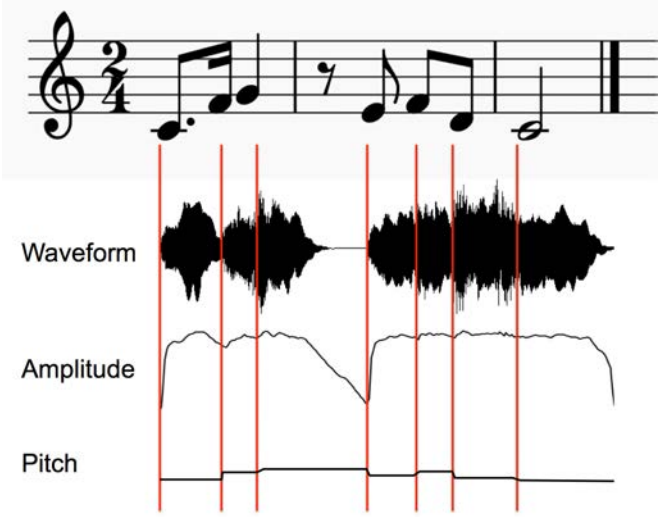

Figure 1 – Amplitude and pitch detection functions for audio onset detection. Vertical lines indicate the note onsets as indicated in the musical notation representation.

5. A Computational and Musician-Friendly Audio Description Scheme

In choosing the audio descriptors that integrate our toolkit, we relied on perceptual criteria for sound description from three musicological theories detailed above: Schaeffer's typo-morphology, Smalley's spectromorphology, and Thoresen's aural sonology—but we did not fully incorporate them into the toolkit because of simplicity, usability, and/or technical reasons. Instead, we selected the criteria that are more adapted for music composition, and whose technical implementations are feasible. A major concern was the use of terminology from music theory and practice to denote the descriptors in the toolkit. Therefore, without disregarding the use of concise concepts, the terms used attempt to facilitate the usability for musicians with a traditional Western music education.

Table 1 organizes the descriptors included in our toolkit according to two principles. The first and topmost organization level splits the descriptors into two categories borrowed from Schaeffer: *matter* and *form*.

The criteria related to matter describe the sound objects' spectrum as a static phenomenon, representing it by a single numerical value, which is meaningful in relation to a finite space constrained by boundaries that represent specific types of sounds. For example, the criteria of noisiness ranges from two typological limits (sinusoidal sounds and white noise), and within these boundaries, sound objects are defined in a continuous scale of real numbers. The form criteria expose the temporal evolution of the matter or the contour of the audio features' evolution and are expressed as lists.

Matter is further divided in two other categories: *main* and *complementary*. While the criteria under the main category provide meaningful descriptions for the totality of sounds

that are audible to humans, the criteria under the complementary category provides meaningful results for limited types of sounds. For example, pitch—a complementary criterion of mass—only provides meaningful results for pitched sounds.

The second categorization adopted in our toolkit is equally borrowed from Schaeffer and correspond to three of his five perceptual criteria for sound description: mass, harmonic timbre, and dynamic. This categorization is used to organize the following sections in which each descriptor is detailed individually. An emphasis will be given to the conceptual basis and musical application of the descriptor in detriment of their mathematical definition, which relies on algorithms from Brent's (2009) timbreID library—to extract low-level audio features from the audio—and an altered [5] version of the Porres's (2011) Dissonance Model Toolbox—to extract sensory dissonance features from the audio.

	MATTER		FORM
	Main	**Complementary**	
Mass	Noisiness	Pitch	Spectral variability
		Fundamental bass	
Harmonic Timbre	Brightness		
	Width		
	Sensory dissonance (roughness)		
	Harmonic Pitch Class Profile		
Dynamic	Loudness		Dynamic profile

Table 1 – Audio descriptors included in our toolkit.

5.1 Criteria of Mass

The mass criteria examine the spectral distribution of a sound object in order to characterize the organization of its components. It not only attempts to detect spectral patterns (e.g., pitch, fundamental bass) but also to provide general consideration of the spectral distribution (e.g., noisiness). The criteria of mass encompass four descriptors: noisiness, pitch, fundamental bass, and spectral variability. The first is a main descriptor of matter, the second and third are complementary descriptors of matter, and the last descriptor falls into the form category.

5.1.1 Noisiness

The noisiness descriptor measures amount of noisy components in the signal as opposed to pitched components. Inspired by Smalley's musicological theory, our measure of noisiness is given by a value that falls in a limited linear continuum, instead of Schaeffer's discrete typology.

The novelty of our descriptor in relation to related research (Ricard, 2004; Peteers and Deruty, 2008) is its computation by a combination of four low-level descriptors—spectral flatness, tonalness, spectral kurtosis, and spectral irregularity—with the aim to provide a better distinction between pitched and noisy sounds. In what follows, we define each low-level descriptor used along with their contribution to the overall computation of noisiness.

Spectral flatness provides an indicator of how noise-like a sound is, as opposed to being tone-like, by computing the ratio of the geometric mean to the arithmetic mean of the spectrum. The output of the descriptor is close to 0 for sinusoidal sounds and 1 for a fully statured spectrum. Within this interval, pitched sounds roughly occupy the interval [0, 0.1], which is clearly poor when compared to that of noisy sounds, which inhabit the rest of the scale.

Tonalness measures the "perceptual clarity of the pitch or pitches evoked by a sonority" (Parncutt & Strasburger, 1994, p. 93) and can be understood as the reverse indicator of the spectral flatness descriptor. However, contrary to the spectral flatness descriptor, it provides a more refined description on the range of pitched sounds as opposed to noisy sounds. The output of the descriptor is high for sounds that evoke a clear perception of pitch, and gradually lower for sounds with increasing inharmonic character.

Spectral kurtosis gives a measure the "peakedness" of the spectrum around its mean value (Peeters, 2004). Spectral kurtosis is particularly good at distinguishing between pitched sounds that range from pure tones (low kurtosis values) to heavy frequency modulations (high kurtosis values).

Spectral irregularity inspects the spectrum from low to high frequencies and denotes how each bin compares to its immediate neighbors (Jensen, 1999). It provides a clear distinction between jagged spectra, i.e. sounds from tones with harmonic or inharmonic spectra jagged (e.g. piano or violin tone) and smooth spectra and spectral distributions formed by "bands" or an array of sounds, which is non-locatable in pitch (e.g. sea sounds and filtered noise).

The combination of descriptors detailed above was heuristically weighted towards a balanced definition between pitched and noisy sounds. The noisiness descriptor ranges between zero and one. Zero represents a full saturated (noisy) spectrum and one represents a pure sinusoid without partials. Within these two extremes the descriptor covers the full range of audible sounds including instrumental, vocal, or environmental sounds.

5.1.2 Pitch

The name of the second descriptor of mass is self-explanatory; it reports the pitch or fundamental frequency of sound segments. Pitch is a secondary criterion of mass, since it only conveys meaningful results for pitched sounds. This descriptor is not contemplated in any sound-based theory discussed in Sections 2 and 3 because it is highly attached to the concept of musical note and does not provide meaningful descriptions for the totality of perceivable sounds. However, the pitch descriptor is adopted here since it may constitute an extremely

important element in the composition process when dealing with pitched audio signals.

Pure Data's built-in object sigmund~ by Puckette is used to compute the fundamental frequency of (monophonic) sounds. The output of the descriptor is in MIDI note numbers.

5.1.3 Fundamental Bass

The fundamental bass descriptor reports the probable fundamental frequency or chord root of a sonority. Similar to the pitch criterion, it is a secondary criterion of mass, because it is constrained to the specific range of pitched sounds. However, contrary to the pitch descriptor it can be applied to polyphonic audio signals. The fundamental bass corresponds to the highest value of the pitch salience profile of the spectrum. The pitch salience of a particular frequency is the probability of perceiving it or the clarity and strength of tone sensation (Porres, 2012). The fundamental bass is expressed in MIDI note numbers.

5.1.4 Spectral Variability

Spectral variability provides a measure of the amount of change in the spectrum of an audio signal. It is computed by the low-level audio descriptor spectral flux (Peeters, 2004), which calculates the Euclidean distance between adjacent spectra. Spectral variability is a form descriptor because it provides a description of the temporal evolution of the sound object's spectrum at regular intervals of 11.6 ms (analysis windows encompass 23.2 ms). The output of this descriptor is threefold: a curve denoting the spectral variability of the sound object, basic statistical values (e.g. maximum and minimum values, mean, standard deviation, and variance) that express characteristics extracted from the aforementioned curve, and finally a single value that expresses the overall spectral variability throughout the sound object (computed by the accumulated difference between analysis windows of 11.6 ms).

5.2 Criteria of Harmonic Timbre

The three musicological theories presented earlier provide little guidance for the formulation of algorithmic strategies to describe the harmonic timbre content of a signal. Schaeffer's criteria of harmonic timbre are very misleading and too inconsistent to be encoded algorithmically. Smalley (1986, 1997) does not provide a specific set of criteria for harmonic timbre; even if he considers this dimension while describing the mass of sound objects under spectral typology. Thoresen's sound spectrum criteria (e.g. spectral brightness) are better adapted for computational use. Additionally, his criteria served as the main inspiration for our work, in particular concerning the adoption of psychoacoustic models of sensory dissonance as harmonic timbre descriptors [6]. The following sections will further detail the four descriptors adopted in our toolkit to characterize harmonic timbre: brightness, width, sensory dissonance (roughness) and harmonic pitch class profile. All harmonic timbre descriptors fall under the main category because they can measure properties of all perceivable sounds, and offer a representation of the units with a single numerical value.

5.2.1 Brightness

The brightness of a sound is correlated to the centroid of its spectrum representation and is expressed by the magnitude of the spectral components of a signal in the high-frequency range (Porres, 2011). Although the root of this descriptor resides in psychoacoustics, one can also find it in Thoresen's (2007) musicological theory, which pinpoints its importance in linguistics—in order to distinguish between the sounds of vowels and consonants—and in music—as a distinguishable factor to perceive different traditional acoustic instruments.

Brightness is computationally expressed by the "center of mass" of the spectrum (Peeters, 2004) in units of Hertz and its range has been limited to the audible range of human hearing, which is roughly from 20 Hz to 20 kHz.

5.2.2 Width

Width [7] expresses the range between the extremities of the spectral components of a sound object. In more empirical terms, we may say that the width characterizes the density, thickness, or richness of the spectrum of a sound.

An exact computational model of the width of the spectral components of a sound poses some problems, because the spectral representation of the audio signal may encompass an amount of uncontrollable noise, even if the ideal conditions during the recording stage were met. Instead of considering a solution for this long approached problem, we adopted a simpler, yet effective workaround: the use of the low-level descriptor spectral spread to measure the dispersion (or amount of variance) of the spectrum around its centroid. In such a way, it does not take into account the extreme frequencies of the spectrum, but rather a significant part of it to express the frequency components' range of a sonority. Like brightness, the output of spectral spread is in units of Hertz.

5.2.3 Sensory Dissonance

The descriptor sensory dissonance models innate aspects of human perception that regulate the "pleasantness" of a sonority. Even if the sensory dissonance is regulated by a few psychoacoustic factors, it is expressed in the current framework by what is considered to be its most prominent factor: auditory roughness. In detail, sensory dissonance describes the beating sensation produced when two frequencies are a critical bandwidth apart, which is approximately one third of an octave in the middle range of human hearing (Terhardt, 1974). The partials of complex tones can also produce a beating sensation when the same conditions are met; that is, when they are a critical bandwidth apart.

5.2.4 Harmonic Pitch Class Profile

The harmonic pitch class profile (HPCP), also commonly referred to as chroma vector (Serrà et al., 2008) [8], is particularly suitable to represent the pitch content of polyphonic music signals by mapping the most significant peaks of the spectral distribution to 12 bins, each denoting a note of the equal-tempered scale (pitch classes). Each bin value represents the relative intensity of a frequency range around a particular pitch class, which results from

accumulating the 25 highest peaks of the spectrum warped to a single octave.

5.3 Criteria of Dynamics

The criteria of dynamics describe the energy of the sound objects in two distinct ways: by a single value that offers a rough representation of its overall loudness, and by a curve denoting the dynamic profile of the unit. The first measure is given by the loudness descriptor and the second representation by the dynamic profile.

5.3.1 Loudness

The loudness descriptor expresses the amplitude of a unit by a single value and is defined by the square root of the sum of the squared sample values, commonly addressed as root-mean-square (RMS). The loudness descriptions are computed by Puckette's object sigmund~, which is included in the software distribution of Pure Data.

5.3.2 Dynamic Profile

The dynamic profile represents the evolution of the sound object's amplitude. It can be useful in cases where the single value of the loudness descriptor is too crude or oversimplifying, such as in the retrieval of sound objects with similar amplitude envelope.

The output of the descriptor is twofold: a curve that indicates the evolution of the energy of the sound object (measured by its RMS at regular intervals of 11.6 ms and analysis window of 23.2 ms) and some basic statistics extracted form the curve, such as minimum, maximum, mean, standard deviation, and variance (see Figure 2 for an example of the dynamic profile and extracted statistics).

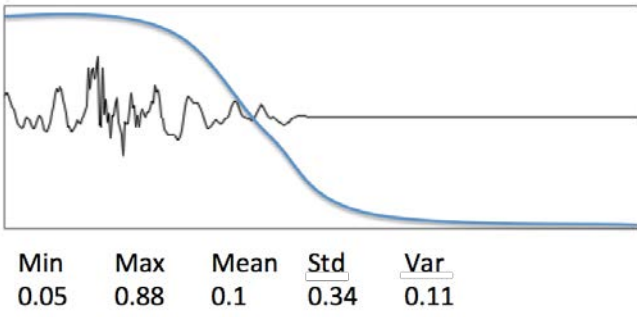

Figure 2 – Dynamic profile of a sound object drawn on top of its waveform. Basic statistics extracted from the curve are presented at the bottom of the Figure.

6. Applications

This toolkit has been developed to extend the terminological and operational level of earGram (Bernardes, Guedes, & Pennycook, 2012), a framework based on a concatenative sound synthesis for computer assisted algorithmic composition that manipulates audio descriptors. For this reason the programming environment of our choice to implement the toolkit was the same as earGram, i.e. Pure Data. The tools can be accessed and used independently of the software earGram though and are accessible at the following address: https://goo.gl/1Pa0KH.

Within the scope of earGram, where these analytical strategies have been evaluated, we were able to apply them to identify and model higher structural levels of the audio data by grouping sound objects into recognizable patterns up to the macro-temporal level. These representations were then used to feed algorithm music strategies commonly applied for the manipulation of symbolic music representations. The resulting framework can be inserted in, and expands upon, the early computational approaches to music analysis-synthesis (Hiller & Isaacson, 1959; Rowe, 1993) on the analysis and automatic generation of music encoded as symbolic representations, towards the use of (digital) audio signals.

We additionally encouraged several composers to explore the toolkit assuming a perspective in which all sonic parameters or criteria for sound description, such as brightness and sensory dissonance, can act as fundamental "building blocks" for composition. This is not to say that every piece composed by these means must use all sonic parameters equally, but that all sonic parameters may be taken into careful consideration when designing a musical work and seen as primary elements of musical structure. For a comprehensive description of some of these works please refer to Bernardes et al., (2012), Gomes et al., (2012), Bernardes (2014) and Beyls, Bernardes, and Caetano (2015).

7. Conclusion

In this paper we presented a Pure Data toolkit for the computational analysis of sound objects' morphology in real-time and offline modes. The analysis tools include methods for segmenting an audio stream into sound objects using MIR strategies for onset detection and beat tracking and a set of audio descriptors that characterize several (morphological) attributes of an audio signal.

The main contributions of this paper are primarily at the *conceptual rather than technical levels*. Our toolkit adopts a reduced number of descriptors in comparison to analogous audio descriptor schemes, selected based on perceptual criteria from phenomenological theories by Schaeffer, Smalley and Thoresen for the analysis of sound-based compositions. By establishing mappings between MIR low-level descriptors and perceptual criteria defined by terms from music theory and practice, we offer a more user-friendly experience for users with a music education background, thus allowing this group to manipulate audio signals through the indirect use of low-level audio descriptors.

A distinctive feature of our toolkit is the adoption of psychoacoustic dissonance models as audio descriptors, which proved to provide robust characterizations of the harmonic timbre of sound objects in generative music applications such as earGram (Bernardes et al, 2013).

Finally, the toolkit shows great possibilities for music composition by offering composers the chance to explore dimensions of sound other than the typical primary pitch and duration attributes of acoustic instrumental and vocal models.

Acknowledgments

This work is financed by the ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-037281 and by the FCT post-doctoral grant SFRH/BPD/88722/2012.

Notes

[1] Shazam, http://www.shazam.com, last access on 16 August 2015.

[2] Moodagent, http://www.moodagent.com, last access on 16 August 2015.

[3] Echo Nest Remix API, http://echonest.github.io/remix, last access on 16 August 2015.

[4] A sound object denotes a basic unit of musical structure analogous to the concept of a note in traditional Western music approaches, but encompassing all perceivable sonic matter (Schaeffer, 1966).

[5] All modifications to the original algorithms were made to improve their computational efficiency.

[6] Please note that Schaeffer clearly rejected the use of psychoacoustic in his solfeggio because (in his opinion) the *in vitro* psychoacoustics experiments did not fully apprehend the multidimensionality qualities of the timbre (Chion, 1983). Given the space constraints of this paper, we cannot describe at length the psychoacoustic dissonance models employed here. To this end, please refer to Parncutt (1989) and Porres (2011).

[7] Please note that Thoresen (2007) adopts a related term, spectral width, to characterize a different characteristic of the spectrum: the mass (called noisiness in our toolkit). Although width can be seen as a "satellite" descriptor of the mass (or noisiness), the two concepts can offer different characterizations of the spectra.

[8] The adoption of the term HPCP instead of chroma vector is due to its widespread use in musical contexts.

Bibliography

Bernardes, G., Peixoto de Pinho, N., Lourenço, S., Guedes, C., Pennycook, B., & Oña, E. (2012). "The Creative Process Behind 'Dialogismos I': Theoretical and Technical Considerations". *Proc. of the ARTECH – 6th Int. Conf. on Digital Arts*, (p. 263-268).

Bernardes, G., Guedes, C., & Pennycook, B. (2013). "EarGram: An Application for Interactive Exploration of Concatenative Sound Synthesis in Pure Data". In M. Aramaki, M. Barthet, R. Kronland-Martinet, & S. Ystad (Eds.) *From sounds to music and emotions*, (p. 110-129). Berlin-Heidelberg: Springer-Verlag.

Bernardes, G. (2014). *Composing Music by Selection: Content-based Algorithmic-assisted Audio Composition*. Ph.D. dissertation, University of Porto, Portugal.

Brent, W. (2009). "A Timbre Analysis and Classification Toolkit for Pure Data". *Proc. of the Int. Computer Music Conf.*, (p. 224-229).

Brent, W. (2011). "A Perceptually Based Onset Detector for Real-time and Offline Audio Parsing". *Proc. of the Int. Computer Music Conf.*, (p. 284-287).

Bello, J. P., Duxbury, C., Davies, M. E., & Sandler, M. B. (2004). "On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain". *IEEE Signal Processing Letters*, vol. 11 (nr 6), p. 553-556.

Beyls, P., Bernardes, G., & Caetano, M. (2015). "The Emergence of Complex Behavior as an Organizational Paradigm for Concatenative Sound Synthesis". *Proc. of the 2rd xCoAx Conf.*, (p.184-199).

Brossier, P. (2006). *Automatic Annotation of Musical Audio for Interactive Applications.* PhD dissertation, Centre for Digital Music, Queen Mary University of London, UK.

Chion, M. (1983). *Guide des objets sonores: Pierre Schaeffer et la Recherche Musicale*. Paris: INA/Buchet-Chastel.

Dixon, S. (2007). "Evaluation of the Audio Beat Tracking System Beatroot". *Journal of New Music Research*, vol 36 (nr 1), p. 39-50.

Gomes, J. A., Peixoto de Pinho, N., Costa, G., Dias, R., Lopes, F., & Barbosa, Á. (2014). "Composing with Soundscapes: An Approach Based on Raw Data Reinterpretation". *Proc. of the 3th xCoAx Conf.*, (p. 260-273).

Gouyon, F., Herrera, P., Gómez, E., Cano, P., Bonada, J., Loscos, A., ... & Serra, X. (2008). "Content Processing of Music Audio Signals". In P. Polotti & D. Rocchesso (Eds.) *Sound to Sense, Sense to Sound: A State of the Art in Sound and Music Computing*, (p. 83-160). Logos Verlag Berlin GmbH.

Grachten, M., Schedl, M., Pohle, T., & Widmer, G. (2009). "The ISMIR Cloud: A Decade of ISMIR Conf.s at Your Fingertips". *Proc. of the Int. Conf. on Music Information Retrieval* (p. 63-68).

Hiller, L. & Isaacson, L. (1959). *Experimental Music: Composition With an Electronic Computer*. New York, NY: McGraw-Hill.

Humphrey, E. J., Turnbull, D., & Collins, T. (2013). "A Brief Review of Creative MIR". *Late-Breaking News and Demos presented at the Int. Conf. on Music Information Retrieval*.

Jensen, K. (1999). *Timbre Models of Musical Sounds*. Doctoral dissertation, Department of Computer Science, University of Copenhagen, Denmark.

Landy, L. (2007). *Understanding the Art of Sound Organization*. Cambridge, MA: The MIT Press.

Mitrovic, D., Zeppelzauer, M., & Eidenberger, H. (2007). "Analysis of the Data Quality of Audio Descriptions of Environmental Sounds". *Journal of Digital Information Management*, vol. 5 (nr 2), p. 48-55.

Parncutt, R. (1989). *Harmony: A Psychoacoustical Approach*. Berlin: Springer-Verlag.

Peeters, G. (2004). A Large Set of Audio Features for Sound Description (Similarity and Classification) in the Cuidado Project. Ircam, Cuidado Project Report.

Peeters, G. & Deruty, E. (2008). "Automatic Morphological Description of Sounds". *Proc. of Acoustics 08* (p. 5783-5788).

Puckette, M. (1996). "Pure Data". *Proc. of the Int. Computer Music Conf.*, (p. 224-227).

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). "The Timbre Toolbox: Extracting Audio Descriptors From Musical Signals". *Journal of Acoustical Society of America*, vol. 130 (nr 5), p. 2902-2916.

Porres, A. (2011). Dissonance model toolbox in Pure Data. *Proc. of the 4th Pure Data Convention*.

Puckette, M., Apel, T., & Zicarelli, D. (1998). "Real-time Audio Analysis Tools for Pd and MSP". *Proc. of Int. Computer Music Conf.*, (p. 109-112).

Ricard, J. (2004). *Towards computational morphological description of sound*. Master thesis, Pompeu Fabra University, Barcelona, Spain.

Rowe, R. (1993). *Interactive Music Systems: Machine Listening and Composing*. Cambridge, MA: The MIT Press.

Schaeffer, P. (1966). *Traité des Objets Musicaux*. Paris: Le Seuil.

Schnell, N., Cifuentes, M. A. S., & Lambert, J. P. (2010). "First Steps in Relaxed Real-time Typo-morphological Audio Analysis/Synthesis". *Proc. of the Sound and Music Computing Conf.*.

Schwarz, D. (2006). "Real-time Corpus-based Concatenative Synthesis with CataRT". *Proc. of the Int. Conf. on Digital Audio Effects*, (p. 279-282).

Serrà, J., Gómez, E., Herrera, P., & Serra, X. (2008). "Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification". *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16 (nr 6), p. 1138-1152.

Smalley, D. (1986). "Spectro-morphology and Structuring Processes". In S. Emmerson (Ed.) *The Language of Electroacoustic Music*, (pp. 61-93). Basingstoke: Macmillan.

Smalley, D. (1997). "Spectromorphology: Explaining Sound-shapes". *Organised Sound*, vol. 2 (nr 2), p. 107-126.

Thoresen, L. (2007). "Spectromorphological Analysis of Sound Objects: An Adaptation of Pierre Schaeffer's Typomorphology". *Organised Sound*, vol. 12 (nr 2), p. 129-14