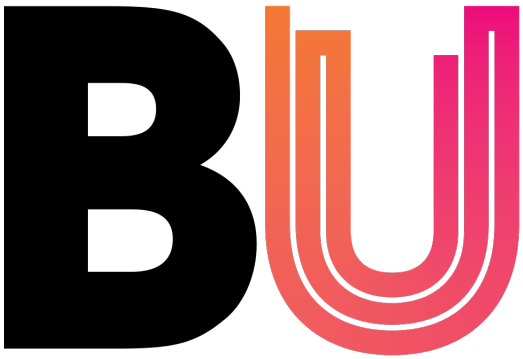
A Perceptually Motivated Approach to Timbre Representation and Visualisation

Sean Soraghan

A dissertation submitted in partial fulllment of the requirements for the degree of Engineering Doctorate



Bournemouth University

Industrial Partner: ROLI
September 2016

Abstract

Musical timbre is a complex phenomenon and is often understood in relation to the separation and comparison of different sound categories. The representation of musical timbre has traditionally consisted of instrumentation category (e.g. violin, piano) and articulation technique (e.g. pizzicato, staccato). Electroacoustic music places more emphasis on timbre variation as musical structure, and has highlighted the need for better, more in-depth forms of representation of musical timbre. Similarly, research from experimental psychology and audio signal analysis has deepened our understanding of the perception, description, and measurement of musical timbre, suggesting the possibility of more exact forms of representation that directly reference low-level descriptors of the audio signal (rather than high-level categories of sound or instrumentation).

Research into the perception of timbre has shown that ratings of similarity between sounds can be used to arrange sounds in an N-dimensional perceptual timbre space, where each dimension relates to a particular axis of differentiation between sounds. Similarly, research into the description of timbre has shown that verbal descriptors can often be clustered into a number of categories, resulting in an N-dimensional semantic timbre space. Importantly, these semantic descriptors are often physical, material, and textural in nature. Audio signal processing techniques can be used to extract numeric descriptors of the spectral and dynamic content of an audio signal. Research has suggested correlations between these audio descriptors and different semantic descriptors and perceptual dimensions in perceptual timbre spaces.

This thesis aims to develop a perceptually motivated approach to timbre representation by making use of correlations between semantic and acoustic descriptors of timbre. User studies are discussed that explored participant preferences for different visual mappings of acoustic timbre features. The results of these studies, together with results from existing research, have been used in the design and development of novel systems for timbre representation. These systems were developed both in the context of digital interfaces for sound design and music production, and in the context of real-time performance and generative audio-reactive visualisation. A generalised approach to perceptual timbre representation is presented and discussed with reference to the experimentation and resulting systems. The use of semantic visual mappings for low-level audio descriptors in the representation of timbre suggests that timbre would be better defined with reference to individual audio features and their variation over time. The experimental user studies and research-led development have highlighted specific techniques and audio-visual mappings that would be very useful to practitioners and researchers in the area of audio analysis and representation.

Contents

List of Figures

List of Tables

Acknowledgments

The passion and dedication that fueled the completion of this research has been instilled by my parents through their continuous support and encouragement of all of my creative endeavours.

The research project was enabled thanks to the financial support of the EPSRC and the Centre for Digital Entertainment (CDE). As a researcher within the CDE I have been afforded unforgettable experiences and lasting friendships.

The creative inspiration for this research project was originated at ROLI, an organisation that has provided untold sustenance in all aspects of my life over the past four years. Here I would list every single employee of the company, past and present, had ROLI's dedication to environmental awareness not had such a profound effect on my habits. Instead I will focus on those with whom I worked most closely and who have accelerated my development as a creative technologist. Chris Fonseka, Vlad Voina, Zsolt Garamvölgyi, Richard Meyer and, in particular, Felix Faire have been inspirational peers and informal mentors in the realm of digital art and creative technology. I would like to thank Roland Lamb and Jean-Baptiste Thiebaut for giving me the opportunity to play a role in such an intellectually and creatively stimulating community.

I have also had the pleasure of collaborating with an independent group of artists known as Analema Group as part of my research, each of whom has been a rich source of creative inspiration. Their names are Oliver Gingrich, Evgenia Emets, Katerina Loschinina, Paulo Ricca, Dario Villanueva, and Marcel Schwittlick.

Finally, I would like to extend my gratitude to my supervisors for the vital roles they played during this research project. My academic supervisor, Alain Renaud, continuously encouraged me to put my research into practice and introduced me to Analema Group, opening up a multitude of opportunities for creative and professional fulfillment. My industrial supervisor, Ben Supper, acted as research mentor at ROLI and ensured a delicate balance was maintained between my academic and industrial pursuits.

Author's Declaration

This thesis contains material that has been previously published. Specifically, section 3.1 contains material that was published in the Proceedings of the International Computer Music Conference (ICMC) in 2014. It also contains material that was presented as a poster at the SIGGRAPH co-located event Computational Aesthetics (CAe) 2015. Section 4.3 contains material that was published in the proceedings of the 2016 New Interfaces for Musical Expression (NIME) conference. Sections 4.3 and 4.6 contain material that was published in the proceedings of the Electronic Visualisation and the Arts conference (EVA) 2016. Also, sections 3.3 and 4.2 contain material that has been submitted and is under review for publication in the Computer Music Journal.

Some of the work presented in this thesis comes from joint research. Specifically, the work discussed in sections 3.2, 3.3 and 4.2 was completed in close collaboration with Felix Faire, a colleague from ROLI, who contributed to the graphical development. The performances discussed in section 4.4 were produced in collaboration with Analema Group – a London-based arts collective. The visualisation system mentioned in section 4.4.3 was developed by Marcel Schwittlick. The visualisation system mentioned in section 4.4.4 was developed in collaboration with Paulo Ricca, and the visualisation system discussed in section 4.5 was developed in collaboration with Paulo Ricca and Dario Villanueva.

Chapter 1

Introduction

Musical timbre is often understood as the 'character' of a sound, the 'sound texture' or 'sound colour'. The American Standards Association defines it as:

> "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."

As discussed in the next chapter, there are a number of problems with this definition, and research from experimental psychology, audio signal analysis and musicology has increased our understanding of the phenomenon. Research into the perception of timbre has shown that it is a *multidimensional* phenomenon and is affected by multiple characteristics of audio. Research into the description of timbre has shown that it is often conceived and described in relation to physical, material and visual qualities. Advances in audio signal analysis techniques have identified quantifiable descriptors of an audio signal that are related to the perceived timbre. Research has also identified correlations between physical acoustic descriptors and perceptual semantic descriptors, suggesting the potential for an improved technical definition that refers to directly quantifiable characteristics of an audio signal. These correlations also suggest the potential for a perceptually-motivated data-driven technique of timbre *representation* where semantic descriptors are used to influence visual mappings for acoustic descriptors.

Electroacoustic music is a type of sound art that situates timbre and timbre variation as the key compositional device (as opposed to pitch and timing). Within the field of electroacoustic music, multiple examples of *graphical scoring* techniques have been developed that constitute methods of perceptual timbre representation from an abstract, subjective (rather than data-driven and objective) perspective.

This thesis presents various novel perceptually motivated data-driven approaches to timbre representation in numerous different contexts. This has been achieved through the experimental development of new techniques, tools and systems for perceptual timbre representation, influenced by suggested correlations between semantic descriptors and acoustic descriptors of timbre from existing studies. The techniques and systems have been put to use in various audio-visual perfor-

mances and installations. These events are described and evaluated with respect to concepts from spectromorphology (a musicological framework for the analysis of electroacoustic music).

The rest of this chapter discusses the motivations behind the research in more detail, and the key problems that are being addressed. The key aims and research questions are then presented, along with a discussion on how they are approached. This is followed by an overview of the rest of the thesis.

1.1 Motivation - The Definition, Perception, and Description of Timbre

1.1.1 Definition

As discussed in the next chapter, the historical definition of timbre has been based around distinct sound categories and their relations, rather than inherent sound characteristics. Since Western tonal composition has historically been concerned first and foremost with pitch, timing and dynamics, the timbre of a sound has traditionally been conceived as the type of sound for which specific sequences of pitches and timings are arranged. The representation of timbre has thus focused on the identification of individual sound types (instrumentation) rather than the individual variation of sound characteristics. Advances in our understanding of the perception of timbre and in the use of timbre variation as a compositional device suggest that a better definition of timbre would refer directly to the spectral content of an audio signal and its variation over time.

1.1.2 Musicology, Perception & Semantic Description

Electroacoustic music often places emphasis on the entire spectral content and its variation over time as one of the main compositional components in music alongside pitch and timing. In the composition and analysis of electroacoustic works, there is a visual and behavioural (motion-related) vocabulary that is drawn upon to describe different aural events, their spectral characteristics, and their spectral variation over time. This visual and behavioural vocabulary is perceptual. The terms describe resulting perceptual impressions that arise in a human listener from the presence of particular spectral properties and processes. This perceptual vocabulary has led to the development of novel timbre notation techniques involving abstract illustrative representations of particular perceptual qualities of timbre. These representation techniques are often subjective and abstract, and are often developed in specific contexts for specific works. Studies investigating the semantic description of timbre in general participant pools (e.g. outside the context of electroacoustic music) have shown that timbre is often conceived in terms of visual, textural and material properties.

1.1.3 Measurement & Acoustic Description

Research efforts from experimental psychology together with sound synthesis techniques from computer music have shown timbre to be a continuous multidimensional phenomenon rather than discrete or singular. Techniques in audio signal analysis and music information retrieval have identified numerous quantifiable audio descriptors that are used to characterise the timbre of the audio. These audio descriptors can be combined to form a high dimensional timbre space that can be explored by varying individual (or multiple) features thereby varying the timbre over time. Novel timbre representation techniques have been developed that are based on the visual representation of these audio descriptors. These representation techniques are low-level and descriptive. Since they represent common audio features, they can be used to represent timbre generally, for any collection of sounds. If these techniques use arbitrary visual mappings for the audio descriptors they can be difficult to comprehend due to their multidimensional nature. The use of perceptually motivated visual mappings for the audio descriptors could produce perceptually salient semantic representation methods for timbre.

1.1.4 Problem Statement - Perceptual Timbre Representation

Although novel forms of timbre representation have been investigated in the areas of musicology and audio analysis, the different techniques have various issues. Current musicological timbre representation methods are often abstract, subjective and context dependent. Current data-driven timbre representation methods often use varied and arbitrary visual mappings.

1.2 Motivation - Interface Design

One can consider the timbre spaces afforded by traditional musical instruments. The mode of exploration through such timbre spaces exists in the different articulation techniques used on the instruments. With digital synthesis engines, there are often high numbers of parameters that can be altered in order to manipulate the timbre. Together, these parameters make up a parameter space, where each parameter may alter the position of the timbre along perceptual dimensions within the timbre space. One of the challenges when designing interfaces for the control of these parameter spaces is in presenting such a complex parameter space in a usable and understandable way.

This section gives a brief description of three different strategies for user interface design, categorising them into the different use cases to which they are suited, namely: *sound design*, *performance*, and *search and retrieval*. The specific interfaces mentioned here by no means constitute an exhaustive list of existing work, but are used to provide examples of the three different interaction paradigms, such that the example systems discussed later can be compared and contrasted in the context of these interaction paradigms.

1.2.1 Sound Design

The term *sound design* is being used in this context to refer to the construction of specific timbres, from the ground up. Most commercial applications for sound design are geared towards this case. The presentation of parameter spaces in such interfaces usually consists of arrays of rotary knobs and sliders and follows the conventions of sound design from an engineering perspective. This type of interface design is rooted in the mimicking of modular synthesis hardware from the Sixties to the Eighties, and has remained the industry standard despite the development of novel synthesis techniques (Seago, Holland, and Mulholland 2004). It is based around what Seago, Holland, and Mulholland (2004) call a 'user specified architecture'. The user patches together sound sources, filters and effects, and then sets the parameters of these various modules in order to construct a required timbre. In the context of timbres paces, this interaction paradigm effectively involves two steps: the construction of the timbre space (linking modules), and the subsequent extraction of a point, or area, within that timbre space (setting parameters). Often what is lacking with interfaces such as these is a strong link between 'task language' (e.g. 'bright,' 'punchy') and 'core language' (e.g. 'filter cutoff', 'envelope attack time') (Seago, Holland, and Mulholland 2004). One way in which this link could be achieved is through the use of a central visualisation of the timbre, based on a perceptually motivated visualisation framework. By using visual, material and textural descriptors of timbre to influence the design of the visualisation, timbre could be represented semantically, with reference to our perceptual understanding and description of the phenomenon. This would facilitate the use of task language in order to refer to timbre-manipulation processes in the semantic, visual domain.

Although these types of interfaces have remained largely constant throughout the past two decades, there have been some examples of innovation. For example, Metasynth by U&I software is a suite of tools for timbre creation and manipulation based on direct interaction with images. The 'image filter,' for example, filters an audio sample by treating individual pixels as tunable stereo bandpass filters where the colour and brightness of the pixel are mapped to stereo position and filter gain, respectively. Similarly, the 'image synth' allows the user to effectively create a sonogram directly. This kind of interface design provides a more low-level approach to sound design than the standard parametric interface, by providing direct access to the spectral characteristics of the sound. Effectively this is a move away from construction of – or exploration through – a parameter space, towards direct manipulation of sound characteristics.

1.2.2 Performance

Performance in this context refers to the real-time manipulation of timbre in a performative setting. Such interfaces are usually built on top of what Seago, Holland, and Mulholland (2004) refer to as a 'fixed architecture'. In other words, what the majority of these interfaces offer is some mode of exploration through an existing underlying timbre space. Indeed, commercial software packages based around user specified architectures often have *simplified counterparts*, where the architecture is fixed, and a collection of specific parameters or macro controls (that alter multiple parameters) is presented to the user for simple – sometimes performative – exploration. Again, interfaces such as these would benefit from a perceptually motivated visualisation that would guide such exploration.

Other examples of performance-oriented interfaces for timbre creation and manipulation offer real-time manipulative control over user specified architectures. For example, reacTable is a tan-

gible interface environment where orientation and proximity of physical modules can be used to alter the synthesis engine's architecture in real time (Jordà et al. 2007). The reacTable system, therefore, offers real-time performative construction and manipulation – as well as exploration – of timbre spaces.

Other performative interfaces make use of different synthesis techniques. One such technique is corpus-based concatenative synthesis (CBCS), which involves the re-synthesis of a target set of sound characteristics (Schwarz 2007). These target characteristics are either explicitly provided, or extracted through audio analysis of a target sound. A corpus of small sound snippets or 'grains' is then referenced to find a set that together can provide a sound with the desired characteristics. CBCS therefore provides access to a high-dimensional timbre space, where the dimensions are defined as the various sound characteristics that are used in the labeling and referencing of sound grains. An important point is that it is not position in the space alone that defines the resulting timbre, but the sound quality of the nearest reference grains in the space. In this way, the full space can be mapped to a lower-dimensional space, which allows for more intuitive exploration (Schwarz 2012). Again, a perceptually motivated visualisation framework would be useful here. It would allow individual grains to be visualised as small sound objects whose physical characteristics (as well as their spatial locations) are indicative of their sound quality. This would increase the information being represented within the low-dimensional space.

The CataRT system is a specific example of a synthesis engine that uses CBCS. It affords the possibility of multiple instruments based 'around the core concept of the timbral sound space that is played by navigation through, and transformation of, its elements' (Schwarz 2012). The kind of performance that is afforded is defined by the interface (either physical or virtual) that is used to navigate the timbre space. Examples of such interfaces include touch surfaces, motion capture systems, and accelerometers.

1.2.3 Search and Retrieval

The third type of interaction is based around the search and retrieval of timbres using some form of user-guided search algorithm (Seago 2013; Gounaropoulos and Johnson 2006). A common method is to use interactive genetic algorithms (IGAs), however these can take a long time to converge on a solution since GAs require multiple generations, and user-evaluation of individual solutions is very time consuming (Takagi 2001). An alternative method is proposed in Seago (2013) called 'weighted centroid localisation' (WCL). This involves a number of candidate solutions that each have a probability that they match the target timbre. At each step, new candidates are created whose positions within the parameter space are equal to the centroid of the existing candidate solutions, weighted by their probabilities of matching the target. The user evaluates this new sound, the centroid is updated, and the process repeats, until the user is satisfied with the generated sound. Effectively, user exploration through a high-dimensional timbre space is guided by continuous user evaluation of reference sounds in comparison to a required target sound. This kind of interface is suited to users without knowledge of sound design but who require specific types of sounds.

A different alternative to IGAs is proposed in Gounaropoulos and Johnson (2006). This involves using machine learning to associate synthesis parameters with timbre labels, thereby providing a perceptually motivated language for timbre description. This technique is described further in section 2.4.2. By constructing robust complex mappings between sound parameters and semantic labels, the process of searching through a timbre space is guided by semantically meaningfully

labeled controls. The drawback to this is that the mappings are hidden within the black-box machine learning algorithm.

1.2.4 Problem Statement - Interface design in commercial tools for timbre creation and manipulation

Despite the innovations discussed in this section, most commercial software interfaces rely on core engineering-focused layouts. The layout and positioning of the individual controls follows the conventions of sound design, beginning with sound generation and following the signal flow through processes such as filtering and effects. Such interface design is suited to the focused creation of individual timbres, but can be limiting to the process of exploration or performative manipulation. It is also restrictive to those users without background knowledge of the sound design process.

1.3 Aims

In response to the two problems outlined in the previous sections, the key aim of this thesis is to develop a *perceptually motivated* approach to timbre representation and visualisation.

Two sub-objectives of the thesis are to examine the potential applications of perceptual timbre representation in both a real-time performance setting and a digital production setting. There are two main research questions being addressed in this thesis. Firstly,

> 'How can correlations between acoustic and semantic qualities of timbre be used to inform the visualisation of timbre in a real-time performance setting?'

Secondly,

> 'How can correlations between acoustic and semantic qualities of timbre be used to inform the design of interfaces for timbre creation and manipulation?'

Quantitative studies have been designed that investigate participant preferences for different visual mappings of timbre features, in different contexts. The results of these experiments, combined with results from other existing studies have influenced the development of novel tools, techniques and systems for perceptual timbre representation. These tools, techniques and systems are used to demonstrate the application of correlations between semantic and acoustic descriptors of timbre in the contexts of live performance and interface design.

In the context of live performance, real-time timbre visualisation systems have been developed and used in collaborative audio-visual performances. These performances have been centred around the relationships between timbre and visual qualities. The timbre visualisation systems make use of the mappings from the quantitative studies, and are analysed and reviewed with respect to concepts from electroacoustic music and the perceptual description of timbre. The performances serve as

technical demonstrations of how correlations between acoustic and semantic qualities of timbre can be put to use in a real-time performance setting. The development of each of the performance-specific systems, as well as the development of a more general system, are used to formalise the key issues in developing real-time perceptual timbre representation tools. This is used in the standardisation and presentation of a generalised approach to perceptual timbre representation.

Similarly, in the context of interface design, tools for timbre creation and manipulation have been developed that make use of semantic visual mappings for timbre features. These novel systems serve as technical demonstrations of how correlations between acoustic and semantic qualities of timbre can be put to use in interface design and development. They provide potential approaches to bridging the gap between core and task language in such interfaces. Again, the development processes behind these novel tools and systems are used in the standardisation and presentation of a generalised approach to perceptual timbre representation.

1.4 Background of ROLI

The previous sections presented the main motivations behind the research presented in this thesis, as well as the key problems being addressed by this research. This section will give some brief background of ROLI and explain how the research aims discussed previously are relevant to the company.

ROLI is a design-led music technology company based in London, founded by Roland Lamb in 2009. The company aims to create novel musical products that broaden the bandwidth of interaction between people and technology. Their flagship product is the Seaboard – a revolutionary digital instrument based on the piano keyboard that introduces five independent dimensions of musical control (X and Y location, pressure, touch velocity and lift velocity). ROLI also aims to develop revolutionary musical software. In 2014 ROLI acquired JUCE – an industry standard C++ framework for audio and GUI development. The JUCE framework was used in the development of Equator – ROLI's digital sound engine and interface.

As a manufacturer of software for musical production and performance, one of the key motivations behind software product development at ROLI is the issue of sound representation and visualisation in digital environments. The experimental procedures and systems described later in this thesis have therefore been developed in collaboration with the ROLI development team. The development of these experimental procedures and systems involved the use of the ROLI Equator synthesiser. Some of the systems described in a later chapter were developed as extensions to the Equator software.

1.4.1 Equator

Equator is ROLI's custom-built sound engine. The Equator interface provides a traditional sound design environment, with multiple sound sources (3 oscillators and two samples), two filters, effects, and multiple modulators. Each module can be enabled or disabled and the routing between modules can be altered. The interface is focused on modulation. The modulator panels (LFOs and envelopes) can be highlighted and then applied to any of the destination parameters (e.g. filter

cut-off value). There are also central panels for touch input information (e.g. pressure, pitch-bend, strike velocity) that allow the touch input to modulate any of the destination parameters.

The Equator interface is very 'traditional' in the sense that the layout follows the conventions of sound design. Sound generation panels are laid out together in one section, filter and effects panels in another, and modulator panels in another. Equator is a professional tool and the interface is very much aimed at experienced sound designers. In terms of the various use cases discussed in section 1.3, Equator is designed for the context of *sound design* (as opposed to performance or search and retrieval). It features what Seago, Holland, and Mulholland (2004) refer to as a user-specified architecture. The Equator interface is shown in figure 1.1.

Figure 1.1: The Equator Interface developed at ROLI.

As discussed previously, one of the major motivations of this thesis is the problem of timbre representation in digital tools for timbre creation and manipulation. Equator is an example of such a digital tool. The layout follows the conventions of sound design and is based around technical parameters that can be controlled using sliders and other control components. The research motivation of interface design is therefore directly applicable to the Equator interface.

Extensive involvement in the development of Equator from the ground up has provided very useful opportunities to experiment with novel extensions to the interface. Two of the systems described in this thesis have been designed as extensions to Equator. They make use of Equator's extensive sound synthesis capabilities and build on top of this novel techniques for sound visualisation and interaction. Equator has not only been used as a base on which to build novel interaction techniques, but also to generate content for experimentation.

1.5 Acoustic Timbre Features

Having discussed the aims and objectives of the research and how they relate to the aims and objectives of ROLI as a company, this section will give specific definition and description of the acoustic properties of timbre that are referred to throughout the thesis. As mentioned when

Table 1.1: Mathematical notation key

Notation	Definition
t_m	The audio signal within time window m
K	The number of real-valued frequency bins
f_k	The centre frequency of bin k
a_k	The magnitude of energy in frequency bin k
p_k	The normalised magnitude of energy in bin k
a_h	The magnitude of energy in the frequency bin that contains frequency h
f_0	An estimation of the fundamental frequency of the signal

outlining the objectives of the research, correlations between acoustic and semantic qualities of timbre are key. The acoustic features of timbre are extracted from an audio signal using signal processing techniques. An understanding of their mathematical definitions and the qualities that they measure is required for discussions of experimental procedures and systems described in later chapters. This section therefore provides an overview of the various acoustic timbre features that are used and referred to throughout the rest of the thesis. Again, this provides the technical background for the methodologies used, experimentation, and systems described later in the thesis.

The timbre representation techniques presented in this thesis focus on real-time reactive visualisations. This ensures that time-critical audio events have direct visual manifestations as the visualisation responds in real-time. As such, the audio features used are focused on the 'instantaneous' description of the frequency spectrum of the audio, for given windows of time. By using features that describe the audio over short windows of fixed-length time, the features can be mapped to visual features such that those visual features are animated in real-time in response to changes in the timbre.

Two types of feature are used to give a description of the audio spectrum: spectral and harmonic. The spectral features describe the structure and shape of the spectral energy distribution of the audio for very short windows of time. The harmonic features are calculated over longer windows of time and describe the harmonic content of the audio.

The choice of features was also motivated by their correlation with semantic descriptors highlighted in previous research. There are certain spectral and harmonic descriptors for which multiple studies have highlighted correlated semantic descriptors of timbre. These suggested correlations are reported along with the explanation of the various timbre features in the rest of this section.

As discussed in chapter 2, an extensive list of acoustic features is provided in Peeters et al. (2011). The features used in this thesis are all taken from that list. The notation in the following mathematical definitions is defined in table 1.1.

Spectral Centroid

The spectral centroid describes the centre of mass of energy within the frequency spectrum, for a given window of time, of an audio signal. It is calculated as:

$$\mu_1(t_m) = \sum_{k=1}^{K} f_k \cdot p_k(t_m). \tag{1.1}$$

Multiple studies have found spectral centroid to be a good indicator of the brightness of a sound (Beauchamp 1982; De Poli and Prandoni 1997; Berthaut, Desainte-Catherine, and Hachet 2010; Zacharakis, Pastiadis, and Reiss 2014).

Spectral Spread

The spectral spread is a measure of the spread of the spectrum around the mean. It is calculated as:

$$\mu_2(t_m) = \left(\sum_{k=1}^{K} (f_k - \mu_1(t_m))^2 \cdot p_k(t_m) \right)^{1/2}. \tag{1.2}$$

Low spectral spread indicates less energy in other areas of the spectrum (not close to the mean) whereas high spectral spread indicates more energy across the spectrum. The notation of μ_1 and μ_2 for spectral centroid and spectral spread, respectively, comes form Peeters et al. (2011).

Spectral Flatness

The spectral flatness is a measure of the noisiness (or conversely the 'pureness') of a signal. Signals with flatness close to 0 will be less noisy. Signals with flatness close to 1 will be close to broadband noise. It is calculated as:

$$SFM_{tm} = \frac{\left(\prod_{k=1}^{K} a_k(t_m) \right)^{1/K}}{\frac{1}{K} \sum_{k=1}^{K} a_k(t_m)}. \tag{1.3}$$

Existing research suggests that the noisiness of a signal is often linked to the 'texture' of the timbre, in semantic description (Giannakis 2001; Berthaut, Desainte-Catherine, and Hachet 2010).

F0 estimation

There are numerous algorithms for the estimation of fundamental frequency (F0). In this thesis the YIN algorithm is used (De Cheveigné and Kawahara 2002), which is based on autocorrelation. The algorithm is simple and efficient, resulting in low latency. This makes it appropriate for the timbre representation techniques described in this thesis, which have real-time constraints.

Harmonic Energy Ratio

The harmonic energy ratio measures the proportion of energy in the spectrum that is accounted for by the first H harmonics. In this implementation, H was set to 15, but it can vary depending on implementation. The harmonic energy is first calculated as

$$E_H(t_m) = \sum_{h=1}^{H} a_h^2(t_m), \tag{1.4}$$

where H =the first H harmonic partials (multiples of F0). H is set to 20, in line with (Peeters et al. 2011). The harmonic energy ratio can then be calculated as

$$HER = E_H(t_m)/E(t_m), \tag{1.5}$$

where $E(t_m)$ is the total energy in the spectrum at time window m. The harmonic energy ratio gives an indication of the level of harmonic energy that is present in the sound, as opposed to noise energy.

Inharmonicity

The inharmonicity measures the deviation of the first P partials from harmonic frequencies (multiples of F0). It is calculated as

$$inharm(t_m) = \frac{2}{f_0(t_m)} \frac{\sum_{h=1}^{H} \left(f_h(t_m) - h f_0(t_m)\right) a_h^2(t_m)}{\sum_{h=1}^{H} a_h^2(t_m)}. \tag{1.6}$$

In this implementation, the peak frequencies found through peak detection in the spectrum are taken as the partials.

The inharmonicity gives an indication of the 'tonality' of the sound.

1.6 Thesis Structure

The remainder of this thesis is structured as follows. Chapter 2 gives a detailed discussion on the related research into the perception, semantic description, acoustic analysis, and representation of timbre. Problems with the traditional definition of timbre are discussed, followed by description and discussion of the main techniques used in the study of the perception and acoustic analysis of timbre. Results of existing studies and their implications are discussed, which provides motivation and background for many of the implementation decisions made in the development of the various systems described later in the thesis. Some of the key existing techniques for timbre representation are discussed and evaluated, both from the perspective of musicology and of acoustic analysis and

audio signal processing.

Chapter 3 documents the implementation of user studies and experimental development aimed at the implementation and evaluation of mapping strategies between acoustic timbre features and semantic timbre descriptors. Such mapping strategies are central to the concept of perceptual timbre representation. They are used to inform the design and development of novel tools, techniques and systems for perceptual timbre representation in the contexts of both performance and interface design.

Chapter 4 discusses the development of these various tools, techniques and systems. The implementation processes and resulting systems serve as technical demonstrations of how perceptual timbre representation can be applied in each context. Chapter 4 also discusses various audio-visual performances that made use of custom-built real-time timbre representation systems. Mapping strategies were developed for each of these performances that were dependent on the technology used, the performance structure, and the performance environment. The research described throughout the thesis was put into practice for these events, and they demonstrate how perceptually motivated timbre representation can be used in a live concert setting.

Chapter 5 brings together the results of experimentation and research-led development of novel systems in order to provide general approaches to perceptual timbre representation that can be applied in various contexts.

Chapter 6 discusses the key findings and draws some key conclusions. The theoretical implications and potential applications are discussed. Finally, limitations of the research are discussed and suggestions for future work are presented.

1.7 Key Contributions & Publications

The preliminary study described in Section 3.1 was published in the proceedings of the International Computer Music Conference (ICMC) 2014. The study makes use of existing audio and verbal descriptors from existing studies, and examines user preferences in the context of 3D visualisations. Results show clear preferences in the case of isolated mappings, but suggest wide variation in preferences when multiple parameters are varied in tandem. The study also makes use of 3D animated visual stimuli which has not been examined in other similar studies.

The representation technique discussed in Sections 3.3.3 and 4.2 is a novel waveform representation technique combining representation of temporal dynamic, timbral, and panning information. A paper has been submitted for publication in the Computer Music Journal and is currently under review. Section 3.3.3 describes an online interactive survey that made use of the novel representation technique and identified significant preferences for specific timbre-feature to visual-feature mappings. Section 4.2 describes the technique in detail and discusses its integration into the ROLI Equator software which involved the use of a novel audio rendering technique referred to as 'offline rendering'.

The real-time audio feature extraction tool described in Section 4.1 is freely available on Github and open for public use. A talk will be given at the Audio Developer Conference (ADC) in London during November 2016, which describes the implementation of this tool using the C++ JUCE

library.

The generative audio-driven visualisation techniques described in sections 4.3 and 4.6 were presented at the 2016 Electronic Visualisation & the Arts London (EVA London) conference, and published in the proceedings.

Section 4.3 describes a novel system for timbre representation that was presented and demonstrated at the 2016 New Interfaces for Musical Expression (NIME) conference and published in the proceedings. It makes use of semantic and acoustic correlations of timbre in the context of complex 3D visualisations that have thus far not been explored in similar systems, combined with motion tracking for gestural control.

Section 4.4 describes a number of audio-visual performances that were produced in collaboration with a London-based arts collective. The performances make use of semantic and acoustic correlations of timbre by mapping acoustic timbre descriptors to parameters of particle systems and fluid simulation algorithms. Most other performances and installations of this kind make use of congruent audio and visual synthesis (as opposed to direct audio-driven visualisations).

Chapter 2

Musical Timbre: Perception, Semantics, Acoustics and Representation

2.1 Introduction

This chapter gives a detailed review of the existing research into the perception of musical timbre. As an introduction, some problems with the definition of timbre are discussed. Motivations and support for an improved definition are then provided. Section 2.2 discusses the technique of multidimensional scaling as applied to the perception of timbre. The concept of a 'timbre space' is introduced and discussed in detail. The concept of timbre spaces is compared with modern parameter spaces found in consumer tools for sound design. It is argued that timbre spaces and their analysis could be put to use in the development of a perceptually-motivated framework for interface design. Following this, section 3.1.2 discusses the study of timbre perception through analysis of its semantic description. It is suggested that robust mappings between semantic descriptors and perceptual dimensions could aid the development of rich visual environments for sound design. Section 2.4 gives an account of some of the key acoustic descriptors that have been proposed for use in the direct measurement and synthesis of timbre. Discussion is also given on how these acoustic descriptors relate to the perception and semantic description of timbre. Mappings between acoustic properties, semantic labels, and perceptual dimensions of timbre are discussed. A review is then given of some existing timbre representation techniques. Finally, some general conclusions are drawn.

2.1.1 Definition

The definition of timbre is a challenging problem. The American Standards Association identifies timbre as '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.' A number of problems with this definition have been identified. Krumhansl (1989) identifies three key issues.

Firstly, it is a negative definition, defining timbre in terms of what it is not. Thus, it leaves no scope for the analysis or measurement of timbre in terms of its inherent features. This problem in particular highlights our lack of understanding of timbre. As Dannenberg (1993) says, 'timbre is by definition that which we cannot explain'.

The second problem is that the definition assumes a complete separation between timbre, pitch and loudness. There is a growing body of research suggesting that the perception of pitch and timbre can interfere in certain contexts. This is discussed in detail later in this chapter. Krumhansl (1989) suggests that much richer information can be obtained from the study of how these three sound attributes interact, rather than the independent study of each.

The third problem with the standard definition of timbre is that the definition originated in the context of traditional musical instruments and is therefore concerned with differences between acoustic sound sources. Digital sound synthesis facilitates the creation of minute changes in timbre that hint at a more direct definition, related to various measurable qualities of a sound rather than its overall characteristic difference from other sounds.

Despite these problems with the classic definition of timbre, and despite 40+ years of research, an entirely comprehensive definition remains out of reach. Efforts have been focused on explaining the differences in perceived timbre with reference to measurable audio descriptors. This has led to the aggregation of a large set of audio descriptors used in the classification, representation and synthesis of timbre. These efforts are discussed later in the chapter.

The remainder of this section will provide the key motivations behind a suggested alternative definition of timbre.

2.1.2 Electroacoustic Music – Timbre as Musical Form

Deutsch (1984) provides a review of early research into the perception of timbre. She points out that many of the important questions have been raised in a musicological setting, which in turn has inspired novel psychological techniques for the study of timbre perception. For example, Schönberg (1922) predicted the composition of sequences of timbres that 'would work with inherent logic, equivalent to the kind of logic which is effective in melodies based on pitch.' Such compositional techniques have been somewhat facilitated with the introduction of digital sound synthesis and computer music (Roads 1996). However, the 'inherent logic' of timbre is less well defined than that of pitch, and decades of musicological discourse and psychological research have focused on its formalisation.

Timbre has historically been linked with instrumentation; Western musical notation denotes timbre simply using articulation techniques (e.g. pizzicato). From the viewpoint of Western tonal music, the process of musical composition is first and foremost one of organising pitches, durations

and dynamics over time. Within this structure, timbre acts on a more macro level, with specific temporally organised sequences of pitches, durations and volumes being assigned to specific timbre categories (i.e. instruments).

Electroacoustic music has since the 1940s been challenging this notion of musical composition. Electroacoustic music is centred around the idea of organised timbre variation as musical structure, and gives less priority to Western tonal pitch classes. As a musical 'style,' it originated from two seemingly opposing schools of musical composition: *Musique Concrète* and *Elektronische Musik*. Musique Concrète was based around the organisation of recordings of natural non-musical-instrument sounds. Elektronische Musik was based around electronic synthesis of sound and direct control of synthesis parameters. What these two schools had in common was the focus on timbre as the main compositional device. As Di Scipio (1994) points out, electroacoustic music presents 'timbre as musical form'. As such, electroacoustic music (whether intentionally or unintentionally) provides an alternative definition of timbre. Rather than 'everything that is not pitch or volume' timbre can be seen simply as

> The spectral distribution of the audio and its evolution over time, as well as the dynamic variation.

In this context, the spectral distribution refers to how the energy is spread throughout the spectrum, and the dynamic variation refers to the variation in overall energy in the signal (i.e. volume). Peeters et al. (2011) also suggest that timbre is better defined in terms of a specific set of characteristics of the audio signal. They suggest a minimum criteria that experiments involving timbre should meet. They propose that measurements of timbre should include information about:

> (1) one measure of the central tendency of time-varying spectral descriptors; (2) one measure of the temporal dispersion of time-varying spectral descriptors; (3) one descriptor of the energetic content of the sound signals and of the temporal envelope of energy; (4) one descriptor of the periodicity (e.g., F0 or noisiness)

2.1.3 Interaction with Pitch and Loudness

Melara and Marks (1990) conducted a study examining interaction between the auditory dimensions of pitch, loudness and timbre. They asked participants to discriminate between tones according to the three different dimensions. There were three main types of task involved: baseline (where only one auditory dimension varied), filtering (where one dimension varied orthogonal to the other) and congruent (where one dimension correlated either positively or negatively with another). In their experiments, Melara and Marks (1990) varied timbre between 'hollow' and 'twangy'. They arbitrarily designated a positive correlation between timbre and loudness as hollow-soft, twangy-loud. They designated a positive correlation between timbre and pitch as twangy-high, hollow-low. They found that there was interference between loudness and timbre, and between pitch and timbre, in the case of filtering. They also found that correlation between pitch and timbre actually improved discrimination performance. Interestingly, positive correlation between loudness and timbre increased discrimination performance whereas negative correlation decreased performance. Melara and Marks argue that these results act as strong evidence to support the claim that timbre, pitch and loudness all interact in some way during auditory perception in humans. They suggest

that all human perception of specific auditory perceptual dimensions (timbre, pitch or loudness) is contextualised by information from the other dimensions.

There are a number of existing studies similar to that of Melara and Marks (1990) that suggest that the perception of timbre and that of pitch are not entirely separable, or at least that they interfere at certain levels in certain contexts (Wapnick and Freeman 1980; Platt and Racine 1985; Beal 1985; Wolpert 1990; Melara and Marks 1990; P. G. Singh and Hirsh 1992; Pitt 1994; Moore, Glasberg, and Proctor 1992; Krumhansl and Iverson 1992; Warrier and Zatorre 2002; Marozeau, Cheveigné, et al. 2003; Marozeau and Cheveigné 2007; Caruso and Balaban 2014).

For example, there is evidence to suggest that timbre variation can sometimes be mistaken for change in pitch (Wapnick and Freeman 1980; Platt and Racine 1985).

In terms of whether pitch variation affects timbre perception, there have been contradictory results. The results of a study conducted by Marozeau, Cheveigné, et al. (2003) suggest that F0 value did *not* affect timbre-dissimilarity ratings between instruments. Conversely, Marozeau and Cheveigné (2007) conducted a study that suggested that timbre ratings were correlated with F0 values. The main difference between the studies is that Marozeau, Cheveigné, et al. (2003) used musical instrument tones and Marozeau and Cheveigné (2007) used synthesised tones.

Results from studies involving acoustic musical instrument stimuli seem to suggest that musical training might have some diminishing effect on the extent to which timbre variation affects pitch perception (Beal 1985; Wolpert 1990; Pitt 1994). There is also evidence that extended tonal context can minimise the effect of timbre variation on pitch perception (Krumhansl and Iverson 1992; Warrier and Zatorre 2002).

Interestingly, there is also evidence suggesting that the ability to perceive timbre contextually is diminished in the presence of varying pitch (Krumhansl and Iverson 1992).

More recently, Caruso and Balaban (2014) conducted a study involving sequentially presented pairs of tones. Participants were asked to compare the tones according to either pitch or timbre. Results of the main experiment suggested that both features interfere with each other, and that the level of variation in the unattended feature affects the degree of interference, which confirms previous results.

2.1.4 Sound Source Identification

Early focus on identification of sounds produced by 'natural' acoustic instruments highlighted the importance of the attack period of a sound, in the identification of timbre (Grey 1975; Wessel 1978; Wedin and Goude 1972; Berger 1964). It has also been demonstrated that the spectral content has a greater effect on the identification of timbre if it changes over time (Saldanha and Corso 1964). A task linked to the identification of timbres is the attribution of multiple sources to various portions of a single audio signal. It is likely a result of an evolutionary process allowing the human ear to group together sounds emanating from shared sources, and similarly separate those originating at different sources (Deutsch 1984). As Deutsch (1984) points out, three affecting factors of this process have been proposed: 'onset synchronicity' (Rasch 1978; Bregman and Pinker 1978), coordinated spectral modulation (McAdams 1982), and harmonicity (De Boer 1976; Mathews and Pierce 1980).

In digital synthesis, the characteristics of timbre are entirely separated from the sound source (i.e. a computer). In such cases, the perception of timbre needs to be studied in isolation. Helmholtz and Ellis (1875) pioneered early research into the perception of timbre by examining the steady-state portion of simple and complex tones. They suggested that simple tones are generally pleasing, complex tones with prominent harmonics are fuller, complex tones with prominent higher harmonics are rougher, and complex tones with only odd harmonics sound less dense.

2.1.5 Redefining Timbre

Electroacoustic music provides an alternative definition of timbre as 'the spectral content and its variation over time'. This definition positions pitch as an individual feature within the overall set of timbre features. Results from experimental psychology show that under certain circumstances variation in pitch can affect the perception of timbre. Similarly, variation in timbre can affect perception of pitch. These results support the alternative definition of timbre as the spectral content of a sound and its variation over time. This definition of timbre also separates it from the task of sound source identification, which makes it an appropriate definition in the context of digital sound synthesis.

2.2 Multidimensional Scaling

The aforementioned problems with the classic definition of timbre have made the study of timbre perception difficult. Since the specific features of timbre are unclear, it is common practice to make use of audio stimuli which are equal in perceived pitch, duration and loudness, in order to ensure that differences between stimuli are judged based on timbre alone. The application of multidimensional scaling techniques has had a major impact on these kinds of studies. This section will give a brief description of the MDS technique. This is used to introduce the concept of perceptual interface design – interfaces for sound design based on perceptual dimensions of timbre.

Multidimensional scaling (MDS) is a technique that has been used for measuring the perceptual distance between timbres. Listeners are asked to make dissimilarity judgments about pairs of audio stimuli, by giving ratings on a scale from 'very dissimilar' to 'very similar'. The terms 'dissimilarity judgments' and 'similarity judgments' will henceforth be used interchangeably. The stimuli are usually equated for perceived pitch, volume and duration. The resulting data from the similarity ratings are then fed into a multidimensional scaling algorithm that positions the audio stimuli in an N-dimensional 'timbre space' where the distance between two stimuli in the space is roughly equivalent to their overall perceptual difference, according to the ratings (McAdams 2012).

One key advantage to the MDS technique, in terms of measuring perception, is that it makes no assumptions about the specific acoustic descriptors of timbre. Rather, it gives an indication of general similarity between individual timbres in a group. When interpreting the resulting timbre spaces, it is possible to make estimations about which audio features specific dimensions of the space relate to. However, these dimensions are perceptual dimensions and aren't generally indicative of specific audio features. This is problematic in terms of timbre definition and description. Extensions can be made to the classical MDS technique that allow the modeling of mapping functions between physical audio parameters and perceptual dimension position. This is discussed in

more detail later in this chapter.

Most MDS techniques are based around minimising an error function of the form

$$err = \sum_{i<j} (d_{ij} - d'_{ij})^2, \tag{2.1}$$

where *err* is the overall error between measured distance ratings from the participants and plotted distances in the timbre space; d_{ij} = the recorded distance between stimuli i and j; and d'_{ij} = the distance between stimuli i and j in an N-dimensional plot (e.g. Euclidean). Kruskal (1964a) calls this the 'raw stress' value. The plotted coordinates for each stimulus can then be estimated using an iterative process. This process involves first estimating the stimuli coordinates, then updating the estimates based on the error value. The number of dimensions in the space (N) is not known a priori, but can be chosen based on analysis of results for different values of N. In the context of timbre perception, this space construction process can be seen as dimensionality reduction. Listeners judge similarities using an unknown number of dimensions of timbre. The MDS technique attempts to highlight the most salient dimensions in a lower-dimensional space.

There are numerous types of MDS algorithms that have been used. They all feature a mathematical formulation of the distance between two stimuli. McAdams, Winsberg, et al. (1995) provide a survey of the various distance functions that have been used. The most simple example is the standard Euclidean distance, defined as

$$d_{ij} = \left(\sum_{r=1}^{R} (x_{ir} - x_{jr})^2 \right)^{1/2}, \tag{2.2}$$

where d_{ij} = the distance between sounds i and j, R = the number of dimensions, and x_{ir} = the position of sound i along dimension r. This is the classical MDS model. It has been used by Shepard (1962a), Shepard (1962b), Kruskal (1964a), and Kruskal (1964b). This model assumes that each listener gives equal importance to each dimension when making similarity judgments.

There are numerous variations on the classical MDS model. For example the INDSCAL model can be used to measure the importance given to each dimension by individual participants (Carroll and Chang 1970). A different variation proposed by Carroll and Chang (1970) can be used to account for specific auditory characteristics of individual sounds. These two variations can then be combined to account for both specific listener weighting of dimensions and specific auditory characteristics in individual sounds (Winsberg and Carroll 1989). The INDSCAL model can be optimised by assuming that each listener belongs to a certain type, or 'latent class,' of listener. This variation is known as the CLASCAL model (Winsberg and De Soete 1993).

2.2.1 Perceptual Interface Design

These techniques for measuring timbre perception shed light on how timbre is perceived both generally (with the classical MDS model) and under specific circumstances (with the various extensions covering attention to specific dimensions, specificities of certain sounds, and specific types of listeners). The use of MDS in the study of timbre perception also highlights the multi-dimensional and continuous nature of timbre. When sounds are arranged as points in a multidimensional timbre

space, it is interesting to imagine what the spaces in between would sound like. In order to explore this, corresponding acoustic audio features are required for each of the perceptual dimensions. Then, as points are moved around inside the space, their acoustic features are parametrically altered depending on their position along each dimension. Wessel (1979) explored this specific idea using a two-dimensional timbre space. One dimension was correlated with the weighted centre of the spectral energy (spectral centre of gravity / spectral centroid / SCG). The other was correlated with the onset attack time. Further details on investigating correlation between perceptual and acoustic dimensions are given later in the chapter. Wessel (1979) developed a simple graphical control interface where each dimension of a 2D space controlled the relevant acoustic parameter. This early prototype example hints at a perceptual approach to interface design, based around perceptually driven exploration of a physical parameter space.

As Wessel pointed out however, at that time there were limited technologies available to take this concept beyond such a simple implementation. In the time since, multiple programming languages have been developed that are dedicated to audio analysis, synthesis, and audio processing (e.g. Max/MSP, Supercollider) as well as libraries and extensions to existing languages (e.g. the JUCE C++ framework). Despite the affordances offered by such technologies, interface design for sound design tools remains largely based around arrays of rotary knobs and sliders, each controlling individual parameters. The exploration of these parameter spaces is guided by engineering knowledge of signal flow and sound processing techniques. This kind of interface design is inspired by the audio synthesis hardware tools that became popular in the 70s and 80s, which featured physical dials and faders and resembled scientific equipment more than expressive control interfaces.

One of the fundamental problems in designing these interfaces is that of high-dimensional representation. The timbre space control interface as suggested by Wessel (1979) would work well in two or three dimensions, but the concept does not scale when trying to explore a perceptual timbre space of four or more dimensions. In such circumstances, the parameter space cannot be directly represented as a physical, explorable space. The go-to method has been to use individual controls for each parameter, as discussed previously.

2.2.2 Conclusion

In summary, the use of multidimensional scaling (MDS) techniques in the derivation of 'timbre spaces' has been one of the most important developments in the study of timbre perception. A timbre space is an n-dimensional space (usually $1 < n < 5$) in which the distance between two points represents the perceptual distance between two timbres (McAdams 2012). The technique involves asking participants to rate pairs of sounds based on whether they sound dissimilar or similar, and then feeding the dissimilarity ratings into an MDS model that positions the sounds in the timbre space. The key advantage of the MDS technique is that it makes no prior assumptions about the specific acoustic properties of timbre - it simply measures perceptual distance. Correlation between acoustic properties and perceptual dimensions can be investigated in order to estimate the acoustic properties that listeners attend to when making similarity judgements. Given such correlations, techniques for generating novel sounds by querying positions in the space can be conceived, which could give rise to perceptually motivated interfaces for sound design. One key issue in the design of such interfaces would be the representation of high-dimensional timbre spaces.

2.3 Semantic Descriptors

Smalley (1994) contrasts the perception of musical timbre in the context of traditional musical instruments with timbre perception in the context of electroacoustic music. The huge variety of timbres in the context of electroacoustic music cannot be easily categorised according to relatable sources. Additionally, for digitally produced sounds, the sound generation is entirely separated from any gesture or action. As Smalley (1994) points out, 'source-cause texture has evaporated'. He goes on to explain what this means for timbre perception:

> Spectromorphological attributes and ideas then evoke and suggest non-sounding substitutes for the extended and dispersed levels, whether these be to do with motion, types of behaviour, spatial experience, energetic phenomena, psychological tensions, and so on.

Research into the verbal description of timbre suggests that one category of these 'non-sounding substitutes' is that of material and textural properties. Numerous studies have been carried out that investigate the different semantic descriptors that are used to describe timbre. There are two key methodologies that have been used in the identification of such verbal axes: 'semantic differential,' and a related technique known as Verbal Attribute Magnitude Estimates (VAME) (Seago 2009).

This section gives descriptions of these related methods and discusses their limitations. The concept of perceptual timbre representation is introduced – the visual representation of timbre using physical, material and textural qualities.

Semantic Differential

The semantic differential method involves participants giving ratings about particular qualities of auditory stimuli, indicated by verbal descriptors. The scales are constructed using opposing verbal descriptors (e.g. dull-bright, rough-smooth). Dimensionality reduction techniques can then be used to extract the most salient axes of verbal description, thus depicting a verbal timbre space (Zacharakis, Pastiadis, Reiss, and Papadelis 2012).

For example, Bismarck (1974) identified four salient verbal axes: *full-empty; dull-sharp; colourful-colourless;* and *compact-diffused.*

In a similar study, Pratt and Doak (1976) identified 3 salient axes of description: *dull-brilliant, cold-warm* and *pure-rich.* It was also found that participants put most emphasis in the *dull-brilliant* scale when discriminating between sounds. However, the axes of description were not completely independent.

This methodology is similar to MDS in that dimensionality reduction is used to extract the most salient axes of description among a large group. Interestingly, most of the studies using this method have identified similar numbers of dimensions to most MDS studies. One drawback of semantic differential is that it can sometimes be difficult to find completely opposing semantic descriptors for each end of a scale. It is also questionable that participants would agree as to the

specific term that should go at the end of a given scale (Kendall and Carterette 1993; Donnadieu 2007).

Verbal Attribute Magnitude Estimates (VAME)

VAME addresses the issue of questionable scales that exists in semantic differential by making use of negations. In VAME, the given scales consist of a verbal descriptor on one end (e.g. *bright*) and its negation on the other (e.g. *not bright*). Kendall and Carterette (1993) conducted a study using wind instrument sounds which contrasted semantic differential with VAME, using the semantic descriptors from Bismarck (1974). They found that, using semantic differential, little differentiation could be made between the stimuli. Using VAME, better differentiation between stimuli was achieved.

Disley, Howard, and Hunt (2006) conducted a study aimed at obtaining a set of semantic descriptors that would be useful in the control of a sound synthesis system. They used the VAME technique, combined with user confidence ratings for each semantic label, for each stimulus. They identified 5 salient axes of description: *bright/thin/harsh*; *bright - dull*; *pure - percussive*; *metallic - wooden*; and *evolving*.

Limitations

One issue with this kind of study is the subjective and context-dependent nature of the semantic descriptors used. The descriptors used will likely vary depending on a listener's culture, listening habits and musical experience. For example, Disley and Howard (2004) showed that US and UK English speakers seem to have different meanings for the terms 'warm' and 'clear' in the context of timbre description.

Studies have sometimes been limited to specific contexts and participant groups in an effort to ensure more homogeneous vocabulary (Nykänen and Johansson 2003; Bellemare and Traube 2005; Prem and Parncutt 2007; Bernays and Traube 2011).

One of the major issues involved in these studies is the choice of semantic descriptors presented to participants. The identification of a list of possible descriptors by the researcher prior to investigation will undoubtedly introduce bias. For the most part, descriptor selections have been based on previous research. Considering the evolutionary nature of language and the frequent introduction of terms into the musical vocabulary, there is a danger that the research may become stagnated and out-of-date if frequent large-scale vocabulary reviews are not implemented. These issues can be somewhat alleviated by allowing listeners to insert their own descriptors and keeping the most commonly used terms (Disley, Howard, and Hunt 2006; Zacharakis, Pastiadis, Papadelis, et al. 2011).

2.3.1 Perceptual Timbre Representation

The majority of semantic descriptors that are highlighted in such studies are physical, material or textural. Acoustic descriptors can be correlated with verbal axes of description, as is the case for

MDS studies and perceptual dimensions. The identification of such correlations would facilitate the high-level description of specific audio features or combinations of features using visual terms. This suggests the possibility of a general visual representation paradigm for musical timbre that makes use of a combination of a perceptual timbre space (from MDS) and a semantic timbre space (from semantic differential studies). Perceptual timbre spaces can be used to highlight the perceptually salient acoustic properties of timbre. Semantic timbre spaces can then be used to identify visual representations for each perceptually salient acoustic property. Such mappings could be used to construct a perceptual visualisation paradigm for timbre. This provides a solution to the problem of high-dimensional representation in the perceptual interface design paradigm outlined earlier in the chapter. Perceptual dimensions could be represented as material, physical, and textural properties of a 'sound object'. The process of sound design would then become a perceptually-driven process of visually 'sculpting' or shaping the desired sound. This relies on the specific mappings between perceptual, semantic, and acoustic properties.

2.3.2 Conclusion

The description and representation of timbre has, in the past, been limited to instrumentation or sound source. Identification of the timbre of a sound is often linked to the type of source that produced the sound. The introduction of digital synthesis and digitally synthesised sounds has shifted the point of reference for certain timbres towards metaphor, rather than specific source attribution. Research into semantic descriptors of timbre using techniques such as semantic differential and VAME has shown that a large proportion of vocabulary used to describe timbre has a visual context, with descriptors often being textural, physical or material terms. This suggests the possibility of a general visualisation paradigm for musical timbre. Such a visualisation paradigm would require robust mappings between acoustic properties that affect the perception of timbre, and semantic labels given to those perceptual dimensions.

2.4 Perception, Semantics and Acoustics

Having discussed the perception and description of timbre in the previous two sections, this section provides descriptions of various physical audio features that have been used in the measurement and analysis of timbre. Discussion is given on their relation to perceptual dimensions (from MDS) and semantic descriptors (from semantic differential studies). Three commonly used techniques for the identification and evaluation of mappings from acoustic features to semantic descriptors are discussed: correlation analysis, user studies, and machine learning. Various issues involved in existing user studies are highlighted, providing reference for the evaluation of user studies described later in the thesis.

A major objective of the past forty years of research into the nature of timbre has been to identify and formalise a group of audio descriptors that go together to describe timbre. Motivation has come from psychology, composition and performance, and music information retrieval (MIR). In psychology, physical descriptors are needed for the various perceptual dimensions that are indicated through experimental research (Grey and Gordon 1978). Composition and performance can be enhanced through direct control of the acoustic descriptors that control timbre (Wessel 1979). In MIR, the acoustic descriptors can be used as training features in the automatic classification of

timbre (Agostini, Longari, and Pollastri 2003), and as target features in content-based search and retrieval of audio (Wold et al. 1996).

This has led to considerable research efforts in these fields. One problem is that the different fields have varying concerns, priorities and constraints. In psychology and perception, the acoustic descriptors must match as closely as possible the perceptual dimensions on which listeners discriminate sounds. In automatic timbre classification, the focus is on plausible clustering of sounds rather than perceptual significance of descriptors. In addition to this, there are often varying methodologies within fields. As Peeters et al. (2011) point out, studies on perception often use varying sound stimuli and description extraction techniques, and machine learning research often features variability in classification methods and feature extraction techniques. It is therefore difficult to ascertain a globally useful set of descriptors that can liberally be applied across more than one particular field. With this in mind, Peeters et al. (2011) provided an extensive list of analysis tools (the Timbre Toolbox for MATLAB). They include four input representations: the temporal energy envelope, the short-time Fourier transform (STFT amplitude and STFT power), the sinusoidal harmonic partials, and the gamma-tone filter-bank model (Patterson, Allerhand, and Giguere 1995). From these they extract 32 audio descriptors. Finally, they perform information redundancy analysis and hierarchical clustering to identify the number of independent groups of descriptors. The audio descriptors are categorised into temporal parameters, spectral parameters, harmonic parameters, and spectro-temporal parameters.

Peeters et al. (2011) subject all of the acoustic descriptors to hierarchical clustering. Approximately ten relatively independent types of descriptor emerge. Two categories relate to the spectral shape, one category relates to the temporal descriptors, and another relates to the noisiness or periodicity. The other categories include fewer descriptors (one or two) including spectral shape descriptors, spectral variation, and amplitude and frequency of temporal energy modulation. In their analysis, Peeters et al. (2011) used a large, heterogeneous dataset (6000+ sounds from the McGill University Master Samples - MUMS - with varied dynamics, durations and pitches) ensuring the generalisability of their results to a very wide range of acoustic instruments. The descriptors are listed in table 2.1.

The list of descriptors provided by Peeters et al. (2011) includes spectral variation as the only spectro-temporal parameter. In the field of human auditory processing, a 2-dimensional FFT of an STFT (spectrogram) has been used to obtain the spectro-temporal modulation power spectrum (MPS) (Chi, Gao, et al. 1999; Chi, Ru, and Shamma 2005; Elliott and Theunissen 2009; N. C. Singh and Theunissen 2003). This modulation spectrum captures both spectral and temporal modulation simultaneously (Elliott, Hamilton, and Theunissen 2013). The x-axis shows the temporal modulation (Hz), and the y-axis shows spectral modulation (cycles/kHz). Therefore, vertical areas of high density indicate particular oscillations around certain frequencies. Horizontal areas of high density indicate widespread temporal variation in a particular frequency band.

Mel-frequency cepstrum coefficients (MFCCs) have also proven very useful in the classification of instrument timbres (Eronen 2003; Heittola, Klapuri, and Virtanen 2009; Chudy and Dixon 2013) and voice timbres (Tsai and Wang 2006).

Table 2.1: Acoustic timbre descriptors from Peeters et al. (2011).

Temporal Parameters	
Autocorrelation coefficients	The autocorrelation coefficients describe the overall periodicity of the signal in the time domain.
Zero-crossing rate	The number of times the value of the signal crosses the zero axis. Noisier signals tend to have higher zero-crossing rates.
Attack estimation	The estimated attack portion of the signal (see Peeters et al. (2011) for details of the estimation algorithm).
Log-attack-time	The base 10 logarithm of the attack time.
Attack slope	The average slope of the attack duration of the signal.
Decrease slope	A representation of the rate of drop in energy over the duration of the signal. Percussive sounds have steeper decrease slopes.
Decay	The length of the decay portion of the signal.
Release	The length of the release portion of the signal.
Temporal centroid	The overall centre of gravity over the whole spectral energy envelope. It can be used to separate percussive and sustained sounds.
Effective duration	Indicates the perceived duration of the signal.
Energy modulation (tremolo)	The energy modulation has an amplitude and frequency value, and is modeled using a sinusoidal curve.
RMS-energy envelope	Estimated amplitude envelope, useful for attack time estimation and transient detection.
Spectral Parameters	
Frame energy	The sum of the squared amplitudes from STFT or sinusoidal deconstruction analysis, for a single frame.
Spectral centroid	The spectral centre of gravity of a single frame.
Spectral spread	The spread (standard deviation) of the spectrum around the mean.
Spectral skewness	Indicates the level of asymmetry in the spectral distribution.
Spectral kurtosis	A measure of the flatness of the spectrum (from flat to single-peaked) around the mean value.
Spectral slope	The result of linear regression performed on the amplitude values of the spectrum.
Spectral decrease	The average of the spectral slopes between frequency fk and frequency $f1$.
Spectral roll-off	The upper bound of the frequency spectrum within which 95% of the energy lies.
Spectral flatness measure	The overall flatness of the spectrum. Noisier signals have a higher spectral flatness.
Spectral crest measure	The result of a comparison between the maximum value and the mean of the spectrum.
Harmonic Parameters	
Harmonic energy	The portion of the signal energy containing harmonic partials.
Noise energy	The portion of the signal energy not containing harmonic partials.
Noisiness	The ratio of noise energy to overall energy.
Tristimulus	The tristimulus coefficients are three coefficients that effectively characterise the initial harmonics of the spectrum.
Fundamental frequency	The base frequency of the spectrum.
Inharmonicity	A measure of difference between partial frequencies and purely harmonic partial frequencies.
Harmonic spectral deviation	The deviation of partials from the averaged energy envelope.
Odd-to-even harmonic energy ratio	The ratio between odd harmonic and even harmonic energy.
Spectro-Temporal Parameters	
Spectral Variation	The variation of the spectral energy over time.

2.4.1 MDS Dimensions

In perceptual studies that make use of MDS, the positions of the different stimuli along a given dimension can be compared with the values of a given acoustic descriptor for each of the stimuli. This can highlight correlations that suggest which acoustic features participants are making judgments on when differentiating timbres (McAdams 2012).

For example, multiple studies have shown high correlation between spectral centre of gravity (SCG or 'spectral centroid') and a particular dimension resulting from MDS analysis (Grey and Gordon 1978; Krimphoff et al. 1994; Kendall, Carterette, and Hajda 1999; McAdams, Winsberg, et al. 1995). Similarly, log-attack time has high correlation with a particular perceptual dimension in some studies (Krimphoff et al. 1994; McAdams, Winsberg, et al. 1995). Spectral fluctuation - the average of the correlations between amplitude spectra in adjacent time windows - showed high correlation with the third dimension in Krimphoff et al. (1994) and McAdams, Winsberg, et al. (1995). However, it only accounted for 29% of the variance along the third dimension in McAdams, Winsberg, et al. (1995).

Caclin et al. (2005) conducted a confirmatory study using synthetic tones, where the log-attack time, spectral centroid, spectral flux (level of variation in spectral energy over time), and spectral fine structure were all varied parametrically. The study confirmed that participants used log-attack time, spectral centroid, and spectral fine structure when making timbre judgments. However, the use of spectral flux in timbre judgments was less clear, and seemed to depend on the number of other parameters that vary concurrently. This is an interesting outcome, and reflects certain results that have come from studies involving the interplay between timbre perception and pitch perception as discussed previously in section 2.1.3. Such results suggest that our perception of timbre can be influenced by the magnitude of change in each of the different timbre dimensions. It may be the case, for example, that if certain dimensions of timbre remain relatively stable, we give greater attention to other dimensions in order to distinguish sounds.

In a study investigating the link between perceived emotion and timbre, Wu, Horner, and Lee (2014) found that participants were using the spectral fine structure (specifically the even/odd harmonic ratio) to a large extent for stimuli with equalised attack time and spectral centroid. They conclude that even/odd harmonic ratio is a strong candidate for the third most salient acoustic property affecting timbre, after SCG and attack time. The use of perceived emotion ratings may not be the most robust form of measurement, particularly when the stimuli are traditional acoustic instruments (as was the case in Wu, Horner, and Lee (2014)) since the emotional ratings could be influenced by participants' previous knowledge or experience of the instruments in different settings.

The technique of examining correlation between acoustic features and perceptual dimensions can be useful for highlighting perceptually salient quantitative descriptors of timbre. However, it is important to remember that the process of constructing a perceptual timbre space is one of dimensionality reduction. The dimensions would therefore likely be better represented using a weighted combination of different acoustic parameters.

2.4.2 Semantic Labeling of Acoustic Features

Identification of correlations between acoustic features and perceptual dimensions has helped in the identification of particular acoustic features that are important in the perception of timbre. Given such a set of acoustic features that can help provide a quantitative representation of timbre, it would be advantageous to obtain salient mappings between the acoustic features and specific semantic descriptors of timbre, such that timbre can be visualised effectively. Indeed, this would provide the groundwork for the perceptually motivated approach to interface design that is being proposed in this thesis.

Creasey (1998) provides an extensive list of semantic labels and their possible acoustic correlates (sourced from various studies). The list is summarised in table 2.2. As Creasey (1998) points out, the major problem with these suggested relations is that the results of different experiments depend on the timbre spaces that were used. This, combined with the fact that the results from the various studies seem to be inconsistent in places, calls into question the generalisability of the results. As well as this, it is likely that each semantic descriptor refers to a weighted combination of various acoustic descriptors. It is also likely that there is overlap of acoustic descriptors between semantic descriptors. These problems make it difficult to describe the complexities of semantic descriptors in a consistent way.

Other approaches to link semantic labels with acoustic features can be grouped into three categories: correlation analysis, user studies, and machine learning.

Correlation Analysis

Attempts have been made to combine MDS with verbal attribute ratings, such that the resulting perceptual and semantic spaces can be compared (Zacharakis 2013). Kendall and Carterette (1993) and Kendall, Carterette, and Hajda (1999) took this approach but found only slight similarity between the spaces. Faure, McAdams, and Nosulenko (1996) used MDS to obtain a perceptual timbre space and combined this with verbal descriptions of the perceptual distances, resulting in 22 semantic descriptors. Most of the descriptors correlated with many perceptual dimensions, making mappings between semantics and perception unclear.

By conducting correlation analysis between acoustic features and semantic descriptors, Zacharakis, Pastiadis, and Reiss (2014) found the following correlations: *harmonic energy distribution – texture*; *spectral centroid variation and inharmonicity – thickness* (related to mass and luminance); *fundamental frequency – mass or luminance* (depending on native language). There is a large consensus for the association between spectral centroid and the perceived 'brightness' of a sound (Beauchamp 1982; De Poli and Prandoni 1997; Schubert, Wolfe, and Tarnopolsky 2004; Schubert and Wolfe 2006).

Alluri and Toiviainen (2010) conducted a study on the perception of polyphonic timbre in which they first identified a set of semantic descriptors using the semantic differential method. After performing factor analysis on the descriptors, they identified that the following three captured the majority of the variance in the data: activity, brightness and fullness. These results reflect the descriptors extracted in many studies on monophonic timbre, suggesting that the perception of timbre has some consistency between monophonic and polyphonic stimuli. Alluri and Toiviainen (2010) then performed regression analysis using acoustic descriptors of timbre. The results sug-

Table 2.2: Acoustic correlates of semantic labels from Creasey (1998).

Semantic Descriptor	**Acoustic Descriptor**
Acute	Increasing frequency of the second formant
Bright	Spectral centroid / Spectral spread
Clang	Inharmonicity
Clear	Number of harmonics / harmonic density / decreased spectral slope
Cutting	Energy in the 6th-10th harmonics
Dull	Spectral centroid / Spectral spread
Fat	Inharmonicity
Full	Low frequency energy
Grating	Inharmonicity in the attack portion
Hard	High frequency energy during attack / quick attack of low frequencies / resonance peak narrowness
Heavy	High energy in the range 150-1200Hz
Hollow	Dominance of odd harmonics in the lower range
Intense	Linked to sensitivity range of the human ear
Jarring	Strong harmonics
Lax	Specifically linked to the first and second formant values for spoken vowels
Nasal	Odd-even harmonic balance
Open	First formant frequency
Penetrating	High energy content in high inharmonic partials
Presence	Specifically linked to high energy around 2kHz
Rich	Number of significant harmonics
Reedy	Spectral rolloff / Emphasis on 7th and 5th harmonics
Resonant	Number of harmonics / Spectral slope / Harmonic density
Rough	'Likely to be affected by pitch, spectral structure and musical circumstance'
Small	First and second formant frequency level
Soft	Few harmonics above the first
Thin	Lack of lower frequency
Warm	Inharmonicity of harmonic partials

gested that 'fuller' sounds are correlated with high energy in the low-end range of the spectrum, 'brighter' sounds were correlated with high zero-crossing rates and less energy in the low-end of the spectrum, and sounds with greater 'activity' correlated with high energy in the high-end of the spectrum, greater high-end/low-end energy ratio, and more even spectral distribution. This last finding is particularly interesting. It suggests that the activity descriptor was linked to sounds that had more energy all-round, as well as a higher proportion of energy in the high end of the spectrum. This can be contrasted with the brightness descriptor, which seemed to be linked to high-end weighted spectra with less energy in the low-end.

The SeaWave project by Ethington and Punch (1994) also made use of correlation analysis to present an interface that was based on manipulating the values of perceptually labeled parameters.

One issue that has been reported with such correlation analysis is that results often show that specific acoustic features or perceptual dimensions (in the case of timbre space analysis) are correlated with *multiple* semantic descriptors (Ethington and Punch 1994; Faure, McAdams, and Nosulenko 1996). This makes the construction of mapping strategies complicated, since it suggests the requirement of *n-m* mappings (as opposed to 1-1) from acoustic feature to semantic descriptor.

User Studies

There is a large body of research investigating the ability of participants to match audio and visual stimuli based on visual mappings for pitch and loudness (see Spence (2011) for a recent review). However, there is less work that has investigated visual mappings for timbre features.

Lipscomb and Kim (2004) conducted a user study that investigated the relationship between auditory and visual features of randomised audio-visual stimuli. As audio features they used pitch, loudness, timbre and duration. The visual features they used were colour, vertical location, size and shape.

One very influential experiment involved the use of nonsense words 'kiki' and 'bouba' and abstract 2D shapes. Results showed that participants paired the 'bouba' word with rounder shapes and 'kiki' with more angular, sharper shapes (Kohler 1929).

More recently, Adeli, Rouat, and Molotchnikoff (2014) conducted an experiment that investigated participants' preferences for different visual representations of various audio stimuli. In comparison to other such studies, this study featured a large participant pool (119 participants) and rigorous in-depth statistical analysis that supports the validity of the results. Results showed a strong correlation between timbre and visual shape. This correlation was not affected by colour or grayscale, suggesting a 'perceptual constancy' in the audio-visual association of timbre and shape. There was also some evidence of associations between timbres and colours, which was not affected by fundamental frequency. 'Soft' timbres were paired with blues, greens, and light grayscales. 'Harsh' timbres were paired with red, yellow and black.

In studies based on the differences between vocalised sounds (Kohler 1929) or distinct instruments (Adeli, Rouat, and Molotchnikoff 2014), the variation between timbre stimuli is not quantified. This can make it difficult to discuss and interpret the results. For example, it often leads to remarks that certain types of timbre (e.g. 'soft') are associated with certain visual qualities. While this is very important in the study of the perception of timbre, the development of

tools would benefit from more exact correlations between individual timbre features and individual visual features.

Giannakis and Smith have carried out a number of studies looking at auditory-visual perceptual correlates (Giannakis and Smith 2000; Giannakis 2001; Giannakis 2006). An investigation into the similarities between visual texture and auditory timbre is presented in Giannakis (2001). The following associations were highlighted: *texture contrast – spectral centroid*; *texture granularity – noisiness*; *texture repetitiveness – distances between pairs of partials.*

Grill and Flexer (2012) conducted an online user study investigating the ability of participants to correctly pair audio stimuli with visual stimuli. For each question there was one audio stimulus and five visual stimuli to choose from. The visual stimuli were generated parametrically using an audio-to-visual mapping strategy. For each question, one visual stimulus was generated using the audio stimulus for that question, and the others were generated using a different random audio stimulus from the set of 100. The visual stimulus chosen by the participant was therefore 'correct' if it was that which was generated from the audio stimulus for that question. The audio stimuli used in the experiment were labeled with ratings for various bipolar verbal descriptors from a previous study (Grill, Flexer, and Cunningham 2011). These verbal descriptors were: *high - low*, *ordered - chaotic*, *smooth - coarse*, *tonal - noisy*, *homogeneous - heterogeneous.* The visual stimuli were directly influenced by the semantics of the verbal descriptors. Inspired by the 'Sound Mosaics' representation method (Giannakis 2001), they consisted of 2D images with repeating grains. The following auditory-visual mappings were used:

- *high - low* – colour brightness and hue
- *ordered - chaotic* – regularity of grains
- *smooth - coarse* – jaggedness of grain outline
- *tonal - noisy* – colour saturation
- *homogeneous - heterogeneous* – variability in colour parameters.

When discussing the results, the authors report the *RMS error*, which is calculated using the number of incorrect associations, weighted by their root-mean-square distance from the correct association. Participants in general had a significantly lower RMS error than would be expected for random choice. A sub-group of *expert-listeners* were identified by the authors. This group of participants had high reported levels of 'expertise in electronic/experimental music' and (reported) 'availability of good listening conditions.' This sub-group of participants showed better performance at the audio-visual pairing task.

One issue with the aforementioned studies is that they have focused on static images. The visual stimuli therefore do not directly represent the evolution of spectral content over time. Another drawback often present in such studies is the lack of a 'control choice'. Experiments often feature a forced-choice design where participants are asked to choose which of a number of contrasting visual representations is most appropriate. It may be the case that none is very appropriate but yet a choice has to be made.

The previously mentioned studies have also focused on 2D visual stimuli. Other research has explored the idea of visualising timbre in 3D space and representing different timbres as virtual

Table 2.3: Suggested acoustic-semantic mappings from existing research

Semantic Descriptor	**Acoustic Feature**
Luminance / Brightness	Spectral Centroid (Beauchamp 1982; De Poli and Prandoni 1997) (Berthaut, Desainte-Catherine, and Hachet 2010) (Zacharakis, Pastiadis, and Reiss 2014)
Texture / Activity	Harmonic Content (Giannakis 2001) Noisiness (Berthaut, Desainte-Catherine, and Hachet 2010) Spectral Centroid (Alluri and Toiviainen 2010)
Mass / Density	Low Frequency Energy (Alluri and Toiviainen 2010) Spectral Centroid (Zacharakis, Pastiadis, and Reiss 2014)

3D objects. For example, the TimbreFields project by Corbett et al. involves the simulation of timbre variations produced by physical objects through physical contact such as touch (Corbett et al. 2007). It is a Virtual Reality project that simulates widely-varying timbres based on user location relative to the object and point of interaction on the object.

In Berthaut, Desainte-Catherine, and Hachet (2010), a user study was conducted in which participants were presented with various audio-visual stimuli and their mapping preferences were measured. The audio features used were pitch, volume, noisiness and spectral centroid. The visual stimuli included animated 3D shapes. Results suggested that participants preferred brightness ('colour lightness') as a visual mapping for spectral centroid, and texture rugosity as a visual mapping for noisiness.

The results of various user studies and regression analysis studies are summarised in table 2.3. A major issue with most existing user studies of this type is that they feature limited participant numbers. This calls into question the generalisability of the results. They are also usually performed in controlled lab environments and rarely implemented 'in the wild,' which raises questions about their validity.

Machine Learning

Machine learning techniques can be used for automatic classification of timbres (see Herrera et al. (2000) for a review). By defining a set of synthesis parameters and using them as inputs to a classifier, a measurement can be obtained of how much a given parameter configuration fits a given timbre label. This involves training the classifier using labeled parameter configurations. An iterative process can then be employed in order to quantify how much each parameter needs to change in order to approach a given timbre label, or semantic descriptor. Such a technique was used by Gounaropoulos and Johnson (2006).

An advantage of using machine learning to extract mappings in this way is that arbitrarily complicated n-m mappings can be obtained, which facilitates the representation of groups of synthesis parameters using global perceptual labels. However, this may also be a drawback, as such mappings will be black-box in nature, and hard to inspect. Automatic mapping strategies are likely to remain linked to the specific system or technique that is used to obtain them. Although there are likely to be similarities in mapping strategies that are obtained using related methods, such

similarities are very difficult to quantify due to the black-box nature of many machine learning methods.

2.4.3 Conclusion

There has been a substantial body of research investigating the acoustic descriptors of timbre. This has come from various fields including psychology, composition and performance, and music information retrieval (MIR). The different fields have different requirements and constraints, which has led to difficulty in collecting a global set of descriptors. Peeters et al. (2011) have contributed towards this goal with their extensive list of descriptors. Acoustic descriptors of timbre are often separated into spectral and temporal features. Research from the field of auditory processing has looked at the spectro-temporal modulation spectrum (MPS), which captures rich spectro-temporal features of a sound or set of sounds.

Acoustic features can be used in regression analysis of timbre space dimensions from MDS studies in order to investigate which acoustic features are correlated with which dimensions. Existing research has suggested spectral centre of gravity (SCG) and attack time as salient correlates, as well as spectral variation and spectral fine structure (e.g. even/odd harmonic ratio). However, the mapping between acoustic features and perceptual dimensions likely involves a weighted combination of different acoustic features for each dimension, since MDS is a process of dimensionality reduction.

Semantic timbre spaces resulting from semantic differential studies can be compared and aligned with perceptual timbre spaces from MDS studies. The results can be difficult to interpret since semantic dimensions often correlate with multiple perceptual dimensions. This again suggests a weighted combination of mappings rather than simple 1-to-1 mappings between semantic labels and perceptual dimensions. User studies can also be used to investigate user preferences for specific visual mappings of acoustic timbre features. There are numerous issues with existing user studies that could be improved, including non-quantifiable timbre differences, the use of static 2D visual stimuli, limited participant numbers, and controlled lab environments. Machine learning can also be used to extract complex mappings from aural to visual properties. However, these mappings are 'black box' in nature, which means that it may be difficult to directly inspect the produced mapping strategies.

2.5 Representation

Having introduced the idea of perceptual timbre representation and its dependence on mappings between acoustic timbre features and semantic timbre descriptors, this section will review some existing representation techniques for musical timbre.

Western tonal music notation represents timbre simply using instrumentation and articulation indication. This is the simplest form of timbre representation and is based on the idea of timbre as sound category and sound source attribution. Electroacoustic music as an art form highlighted the need for a more detailed form of representation of timbre in a musical context. This, combined with the introduction of the digital computer and digital sound production practices has influenced the

research and development of numerous techniques and methods for the detailed representation of timbre and its variation over time. This section will discuss some of these techniques and methods. Firstly, a discussion is given on the concept of 'spectromorphology' – a framework for the analysis and description of electroacoustic music. This is followed by a consideration of graphical scores in the context of electroacoustic music, which leads into a discussion of more generally applied techniques of data-driven representation of temporal timbre variation. Finally some techniques for real-time audio-reactive timbre visualisation are reviewed.

2.5.1 Spectromorphology

Attempts have been made to formalise and categorise the compositional concepts involved in electroacoustic music in the same way Western musical notation formalises the compositional concepts of Western tonal music. In particular, Smalley's (1997) introduction of the concept of 'spectromorphology' has been very influential. Spectromorphology refers to spectral content ('spectro-') and how it evolves and changes over time ('morphology'). Smalley provides a very detailed framework by which electroacoustic compositions can be analysed with respect to these properties.

A brief review of some of the various concepts outlined by Smalley (1997) will now be given. Smalley's discussion of these concepts is quite artistic in nature and is predominantly concerned with the appreciation and analysis of electroacoustic music. However, some parallels can be drawn between the concepts of spectromorphology and results from psychological experiments into the perception of timbre. These concepts are summarised here and compared with different findings about the perception of timbre. This provides context for the critical evaluation of systems and performances described later in the thesis.

Gesture & Texture

Smalley (1997) distinguishes between sound gestures and sound textures. In Western tonal music, the simplest gesture would be a note. Notes can then be grouped together into gestural phrases. An important factor of a sound gesture is that it is perceived as a unit; a self-contained sound event that usually occurs during a sequence of similar gestures. Gestures, therefore, serve to move time forwards and are often presented terms of their similarity or difference to other gestures in a sequence. Gestures can have direct relation to the physical world, or can be completely abstract – entirely removed from any form of physical sound exertion. Gestures, then, are related to the notion of source-attribution in the perception of timbre. Timbres can be experienced in direct relation to physical sound sources and also completely in isolation, without any reference to sound sources. It is important to note, however, that even when a gesture is abstract and has no discernible physical source, the amplitude envelope nonetheless evokes certain physical characteristics of motion. When a gesture is stretched beyond any resemblance of human physicality, they cross into the domain of textures.

Textures arise when any or all of the different sections of the amplitude envelope (attack, decay, sustain or release) are extended beyond 'realistic' expectations. Again this can be related to the perception of timbre. Research has shown that the amplitude envelope is an important factor in the identification of timbres of musical instruments (Saldanha and Corso 1964). It has

also been demonstrated that timbre *identification* happens over very short timescales and takes precedence over pitch identification (Tervaniemi, Winkler, et al. 1997; Robinson and Patterson 1995). Spectromorphological 'textures', then, are concerned with the prolonged evolution of timbral features after the initial expectation has been established (identification).

Where as gestures are presented as individual events that might form sequences, moving time a long linearly, textures are presented over much longer timescales and focus on what Smalley calls the 'internal activity'. Textures are concerned not with the relations between sequential sound events, but with the variation of timbral features over a period of time. When variations in the pitch and dynamics are temporally stretched beyond a certain point, we begin to focus more on the variations in spectral features of the sound. Again there are parallels in the experimental study of the perception of timbre. Results suggest that we focus more on timbre when there is less tonal context (i.e. less variation in pitch) (Krumhansl and Iverson 1992; Warrier and Zatorre 2002).

Structure

In electroacoustic works, there is no form of generalisable structure or hierarchy that can be applied across multiple works. This is in contrast to Western tonal music, where the structure is defined by the musical note as the atomic building block, and notes of different durations building up the hierarchical rhythmic structure. The lack of generalisable structure in electroacoustic works can be explained in the context of the perception of timbre. As mentioned previously in section 2.1.2, electroacoustic music is based around the temporal variation of timbre as the core compositional device. Research from experimental psychology has suggested that timbre is perceived on a more 'absolute' basis than pitch, which is perceived in relative terms (McAdams and Cunible 1992; Krumhansl and Iverson 1992). This would suggest that there is no analogy to the transposition of pitch in the case of timbre, making abstract structural organisation and transformation impossible.

Motions & Behaviours

Smalley categorises different textures and gestures found in electroacoustic music into various types of 'motion and growth processes'. Motions within textures are grouped into four types. 'Streaming' refers to textures occuring simultaneously at different levels in the spectrum. 'Flocking' refers to the collective temporal variation of multiple micro-objects that go together to produce a given aural effect. 'Convolution' and 'Turbulence' refer to chaotic textures within the spectrum. 'Behaviours' are concerned with the relations and interactions between different gestures, textures and motions. Behaviours are categorised in terms of coordination (are aural events synchronised over time?), time dependence (does one aural event seem to 'cause' another?), and coexistence (does one aural event 'mask' the other in the spectrum?). In terms of timbre perception, these motions and behaviours are related to the perception of polyphonic timbres and the interactions between different timbres.

2.5.2 Graphical Scores & Spectromorphological Representation

While Smalley (1997) acknowledges the usefulness of sketching and diagrams for electroacustic composition and analysis, he doubts the feasibility of a general-purpose scoring technique or rep-

resentation system in his original discussion of the concept of spectromorphology. Sonograms (the direct graphing of spectral content over time), Smalley argues, are too direct and 'too objective' since the process of listening to (or composing) electroacoustic music is one of paying (or drawing) attention to specific features of the spectral content. He proposes that there is 'no objective method of achieving a visual spectromorphological representation.' Indeed, Smalley points out that one of the main *advantages* of electroacoustic music over western tonal music is the difficulty in capturing its essence in a notation system. Electroacoustic music highlights the importance of temporal evolution and flux of spectral content and, in Smalley's words, 'writing freezes the experience of temporal flux'.

Despite the lack of a general representation paradigm, graphical scores are an important part of electroacoustic music. Often developed in specific contexts for specific compositions, they are of particular interest to musicologists attempting to transcribe electroacoustic works (Battier 2015). As Alexander-Adams (2015) points out, graphical scores are 'unified through diversity rather than conformity' and they are often considered vital parts of the electroacoustic work to which they refer (Battier 2015; Alexander-Adams 2015).

Graphical scores can be created for a number of purposes. They may serve as documentation of the creative process and general composition notes (see Karlheinz Stockhausen's 1966 work 'Kontakte'). They may be intended as directive guidelines (see Cornelius Cardew's 1967 work 'Treatise') or formal prescriptive instructions for performance (see Franco Evangelisti's 1957 work 'Incontri di fasce sonore').

They may also be created in reaction to existing works as transcriptions, in support of analysis and appreciation (for example the 'intuitive score' created in reaction to Franco Evangelisti's 'Incontri di fasce sonore' by Giorni and Ligabue (1998)). One thing that a lot of graphical scores have in common is the use of shape, form, and texture to represent the evolution of sound parameters over time. Although not an example of an electroacoustic work, a well known example is 'Metastasis' by Iannis Xenakis. The graphical elements in the score include layered linear slopes with varying gradients indicating glissandi to be played by string instruments, which collectively form parabolic curves, as shown in Figure 2.1.

Notating Timbre

Although graphical scores are often diverse and context specific, attempts have nonetheless been made to develop unified graphical notation systems. Considering the importance that electroacoustic music places on timbre and temporal variation of timbre, these notation systems can in effect be seen as 'timbre notation systems'. The concept of a timbre 'notation' system for electroacoustic music is slightly contradictory however, since electroacoustic music goes far beyond the categorisation of distinct timbres into separate events or 'notes' to be notated. Rather, electroacoustic music explores the (timbre-) spaces within and between these events and the process of movement and motion within these spaces. Nonetheless it is useful to talk of timbre *notation* since this grounds the representation systems in the context of timbre variation as musical structure in composition and performance. Here, timbre 'notation' refers to symbolic systems of timbre representation in the context of electroacoustic music. As well as denoting sound events and their duration over time, such representation systems are concerned with charting the evolution and variation of the sound event during its lifetime.

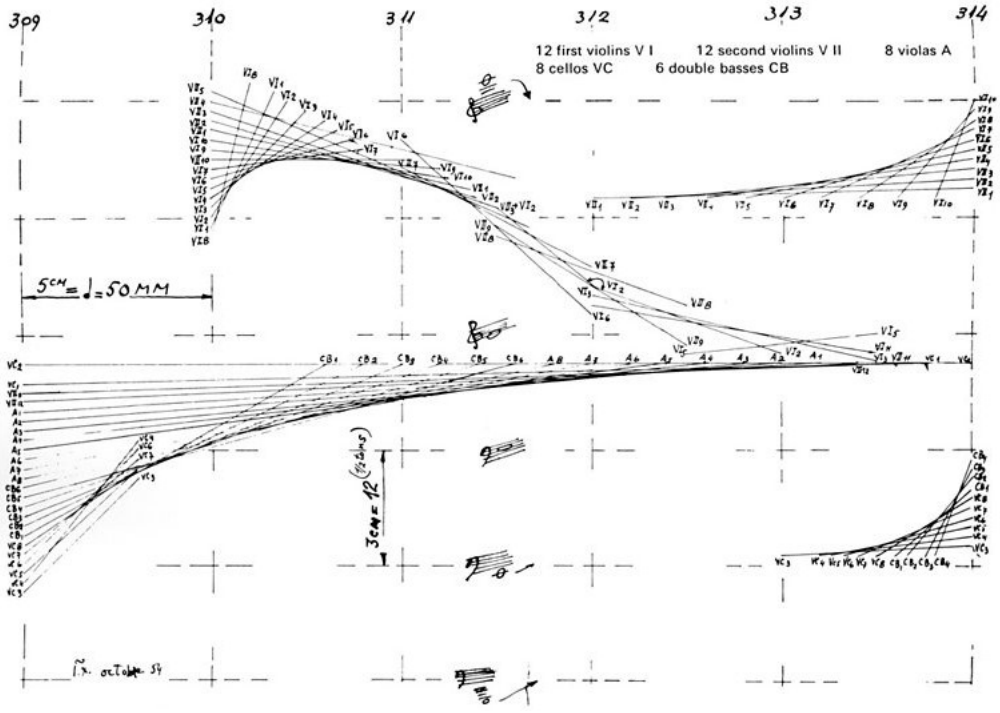


Figure 2.1: A page from the Metastasis score by Iannis Xenakis.

As such, most examples of such representation are perceptually motivated (Battier 2015) and are based on the concepts of spectromorphology from Smalley (1997) or a precursor from Thoresen and Hedman (2007) known as 'typo-morphology'.

For example, Patton (2007) proposed a framework for the notation of interactive electroacoustic music (music involving the interaction between a human performer and a digital system) that he termed 'Morphological Notation' (MN). The idea of MN is based on a 3-dimensional plot where the X-axis represents time, the Y-axis represents continuous pitch, and the Z-axis represents a continuum from noise to tonality, as shown in figure 2.2.

As opposed to tonal pitch notation, the Y-axis is not separated into discrete steps but rather represents a continuous range. On the Z-axis, near values represent noise and far values represent tonality. Thus, events in MN would be represented as 3-dimensional surfaces where the width along the Z-axis indicates the 'range of variance of spectral activity', the width along the Y-axis indicates the variation in perceived pitch, and the width along the X-axis represents the duration. Smalley's discussion of spectromorphology distinguishes between sound *gestures* and sound *textures*. Gestures occur over shorter time frames and are related to physical, natural forms of sound production. Textures extend beyond naturally recognisable sound characteristics and occur over longer time frames. Similarly, Smalley writes of motions, growth processes and behaviours that define the categorisation of – and interactions between – gestures and textures. A representation system like MN would provide visual representation of such concepts through the use of more low-level visual mappings from which the more conceptual categorisations could be drawn.

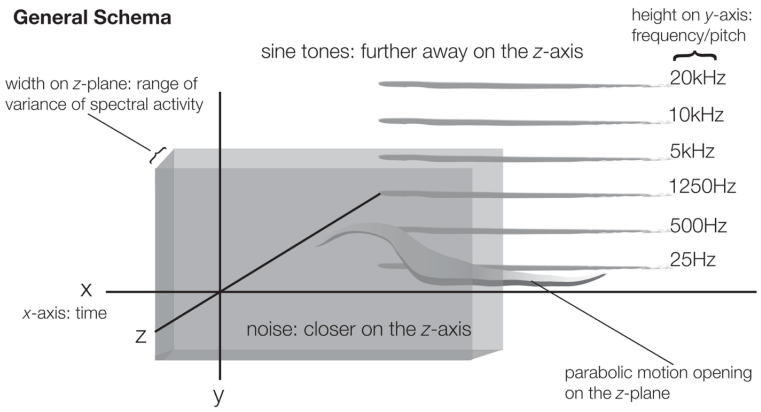


Figure 2.2: 'Morphological Notation' from Patton (2007).

A more abstract, symbolic framework is presented by M. Blackburn (2011), which is based around 'spectromorphological sound shapes'. These are basically graphical, symbolic representations of the various gestures, textures, motions and behaviours defined in spectromorphology. M. Blackburn (2011) discusses the process of 'visual sound sculpting', which involves the use of simple shapes to represent the emergence, temporal evolution, and disappearance of individual sound events. These shapes can then be assembled together into 'sound units'. Sound units can be arranged together to convey the temporal behaviour and interaction between sound events over time. The shape, texture and form of individual spectromorphological sound shapes are directly influenced by the vocabulary of spectromorphology. For example the spectromorphological concept of *agglomeration* is illustrated using multiple compound visual textures indicating parallel motion of different spectral textures. Spectromorphological sound shapes are categorised into *composite* shapes (representing multiple layered sounds) and *singular* shapes (representing isolated sounds). Composite shapes are further categorised into *micro-composite* shapes (for short durations) and *macro-composite* shapes (for longer durations). Whereas the kind of representation paradigm proposed by Patton (2007) is based on the emergence of conceptual categorisations from universal visual mappings, the 'spectromorphological sound shapes' proposed by M. Blackburn (2011) are specifically designed to represent specific concepts. They are intended as creative stimulations for the compositional process and are intentionally left open to interpretation and extension. Representation paradigms such as MN, then, are *objective* and data-driven whereas paradigms such as spectromorphological sound shapes are *subjective* and symbolic. Data driven representations are useful for analysis, detailed performance guidance and unified representation. Symbolic representations are stimulative and useful for both creative inspiration and aesthetic appreciation.

Recently, a research project has been initiated in Malaysia that is dedicated to the development of a timbre notation system based on spectrography. A. Blackburn and Penny (2015) give details of three sub-projects that are involved. One is an ethnomusicology project based around the spectral analysis of performances on Malaysian percussive instruments. The project explores whether spectral analysis can be used to provide visual representations of certain playing styles that are traditionally taught through mnemonic communication. The second project is based around the exploration of extensions to the relationship between performer and score through the use of spectrographs as scores. Questions being addressed include whether spectrographs can be used effectively as forms of timbral notation, and the level of subjective interpretation that can be used

in resulting performances. The third project explores the use of spectrograms in the development of a notation system for timbre in the context of electroacoustic composition, with an emphasis on the documentation and storage of electroacoustic works without reference to the equipment used in their production.

While the timbre notation systems discussed thus far have been concerned with the creation of scores and the composition of works, the third sub-project described by A. Blackburn and Penny (2015) is related to the *transcription* of timbre. The task of timbre transcription has received a lot of attention, particularly in the form of software applications for the automatic analysis and representation of timbral structure in existing works.

For example two software applications for electroacoustic music analysis were presented at the twelfth electroacoustic music studies network conference in 2014 where analysis was a major theme. These were EAnalysis, written by Pierre Couprie at De Montfort University, and a suite of tools referred to as Tools for Interactive Aural Analysis (TIAALS) developed at the University of Huddersfield (Stevenson 2014).

The automatic analysis and transcription of audio data, for whatever purpose, ultimately requires a machine-readable encoding of the audio data such that low-level audio features can be tracked over time in order to segment higher-level features relevant to the particular type of transcription or analysis that is being performed. Significant work has been carried out in this regard by Malt and Jourdan (2011), in the context of analysis of electroacoustic music. They present a review and comparison of the use of three spectral descriptors for the automatic detection of aural events. The descriptors are spectral spread, spectral centroid, and a compound function calculated using the squared product of the spectral centroid and spectral spread, gated by the root-mean-square amplitude. These descriptors are used to detect onsets of aural events with specific spectral characteristics. They conclude with the presentation of a context-specific approach to aural event detection involving: 1) the identification of spectral characteristics of the target event, 2) the identification of appropriate spectral descriptors that accurately represent these characteristics, and 3) the development of an event detection function based on these identified descriptors. This approach to aural event detection contrasts with many other techniques that are designed to be generally applicable across multiple types of event.

2.5.3 Temporal Data Representation

The preceding discussion on timbre notation focused mainly on systems that were developed in the context of musicology, for the purposes of composition, performance and analysis of electroacoustic works. Methods for the representation of timbre and temporal timbral variation have also been developed from the perspective of general low-level audio analysis and sound production, without reference to higher-level compositional structures or concepts. An analysis of some of these techniques will now be given.

BStD

The first of these more generalised timbre representation techniques was admittedly developed with analysis of electronic music in mind and is demonstrated through the analysis of various

electronic works. However, it is included in this discussion of general techniques as it is clearly generally applicable across audio analysis and sound production. BStD is a graphical representation technique for the spectral content of audio and its variation over time base on the Brightness (spectral centroid) and Standard Deviation (spectral spread) of the spectrum (Malt and Jourdan 2011). In the BStD representation, time is represented linearly along the X-axis while spectral spread and spectral centroid are represented using the Y-axis. Vertical strips of colour are rendered such that the mid-point represents the spectral centroid and the height represents the spectral spread. Finally, the intensity of the colour represents amplitude. Figure 2.3 shows an example BStD representation.

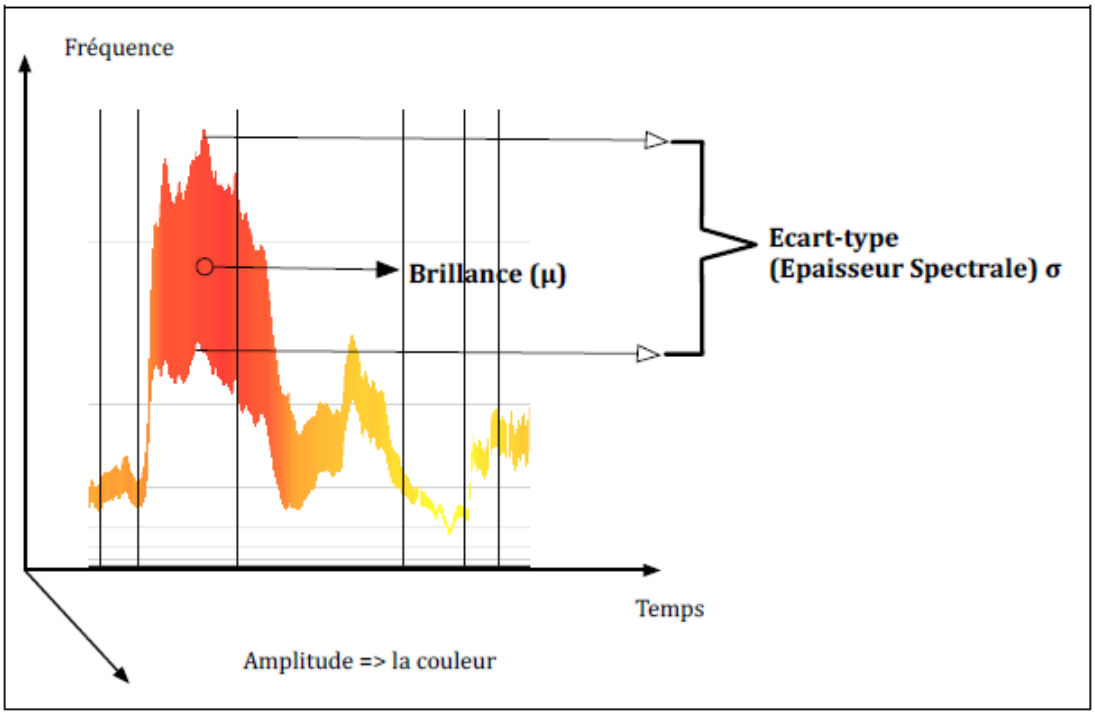


Figure 2.3: An example of the 'BStD' representation from Malt and Jourdan (2011).

The BStD representation is very direct and descriptive. It is similar to a direct spectrogram in that it plots spectral content (Y-axis) against time (X-axis). However it is one step removed from a spectrograph in that it plots general *descriptors* rather than all of the raw spectral data. Other techniques take a similar approach but are further removed in that they use more reduced visual representations of the raw spectral data.

TimbreGrams

Tzanetakis and P. R. Cook (2000) present a method for representing timbre evolution over time that they refer to as a 'TimbreGram'. TimbreGrams use principal components analysis (PCA) in order to cluster similar sounding timbres together and assign colours to timbre groups. TimbreGrams consist of vertical coloured strips that represent short-time feature vectors, as shown in figure 2.4.

PCA is applied to the overall collection of feature vectors and the first three principal components are mapped directly to colour in either red-green-blue (RGB) or hue-saturation-value (HSV) colour space. Tzanetakis and P. R. Cook (2000) found that mapping the principal components into the RGB space produced more 'aesthetically pleasing' results but that using the HSV space provided more clearly defined boundaries.

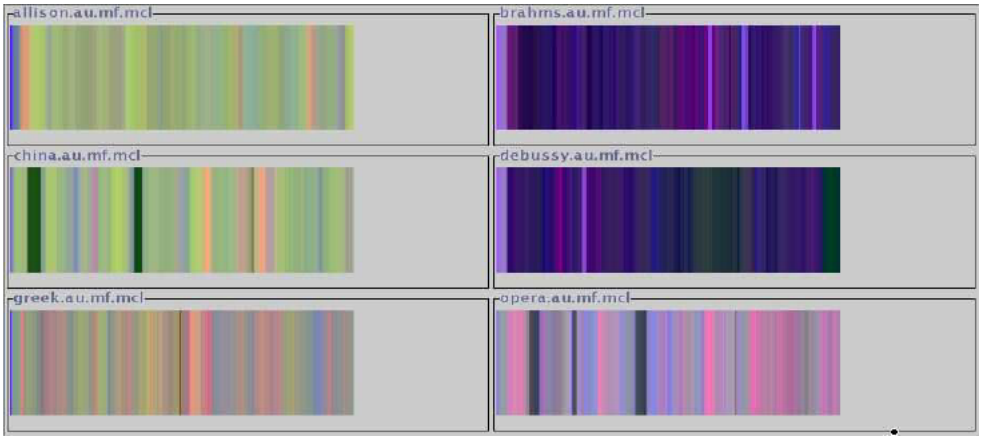


Figure 2.4: Examples of 'Timbregram' representations from Tzanetakis and P. R. Cook (2000).

TimbreGrams can be used in the comparison and distinction of high-level audio qualities. For example, the authors show how TimbreGrams can be used to distinguish between classical music and speech. One of the main drawbacks to TimbreGrams is that they cannot be generated on-the-fly. They require a large data-set in order to highlight existing differences in the data-set. A related drawback is the fact that applying the PCA analysis to different data-sets could result in different colour mappings. This would cause inconsistency and would require adaptation on the part of the user, from data-set to data-set.

Comparisonics Coloured Waveform Display

Rice (2005) presents a method for colouring an amplitude waveform based on frequency content. The method is similar to that of Tzanetakis and P. R. Cook (2000) in that it involves the analysis of adjacent short-time windows. The results of the analysis are then used to colour vertical strips in the resulting waveform image. In the Comparisonics coloured waveform method, low-to-mid frequencies are mapped to blue, mid-to-high frequencies are mapped to green, and high frequencies are mapped to red. Figure 2.5 shows an example of the Comparisonics coloured waveform representation.

The Comparisonics waveform display technique presented by Rice (2005) is a very effective way of presenting rich frequency information about an audio signal and how it varies over time. It improves upon the standard waveform view since it combines both dynamics and frequency data. The method is particularly useful in the context of search and retrieval of audio clips, as well as region isolation. It is now being used in commercial software applications for sound production such as Magix Samplitude and Native Instruments Traktor.

There are, however, areas in which the coloured amplitude waveform technique could be ex-

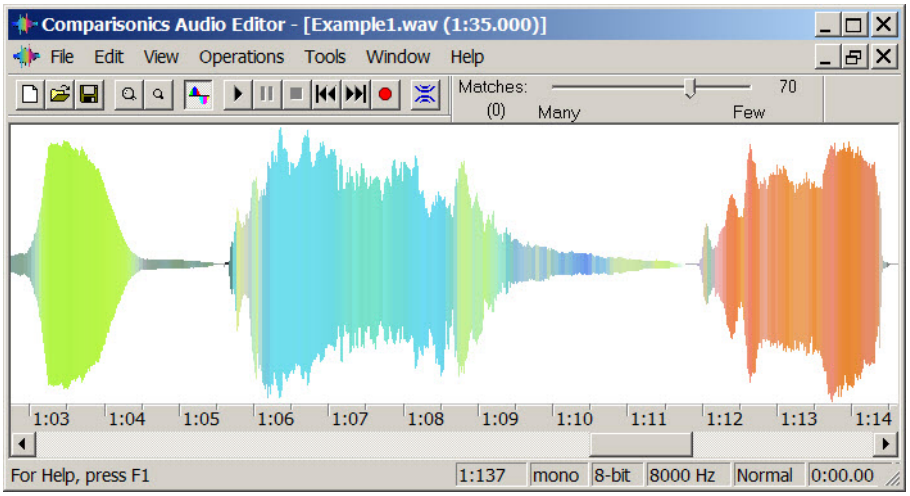


Figure 2.5: An example of the Comparisonics coloured waveform display presented in Rice (2005).

panded. The justification given for the audio-visual mappings is fairly arbitrary. For example, high frequency content is mapped to red in order to invoke a 'connotation of alarm.' Low frequency content is mapped to dark colours in order to invoke an 'ominous' feeling. These mappings are based on emotive qualities rather than descriptive ones. Mappings based on emotive qualities may be less generalisable than those based on direct descriptive qualities. For example it is not always the case that high frequency energy in a signal will invoke a reaction of alarm. Similarly, low frequency energy will not always sound ominous. More robust visual mappings may be identified by examining specific semantic descriptors of timbre, and their acoustic correlates.

2.5.4 Real-Time Representation

Real-time audio visualisation methods can be broadly categorised into two types. Firstly there are methods for scientific visualisation and analysis. For example, real-time FFT visualisers are often used in audio effects plugins for analysis and visual reference. Secondly, there are many media applications that feature real-time visualisers as audio media accompaniment (e.g. the iTunes visualiser). However, in the area of real-time performance and art installations, there has been some interesting work using particle systems and fluid simulation algorithms for audio representation and congruent audio-visual media performance.

Particle & Fluid Simulation

Particle systems and fluid simulations are used extensively in visual media in order to simulate natural phenomena (Stam 2003; Reeves 1983). They are very useful in modeling the behaviour of 'fuzzy' systems such as fire, water or smoke. Particle systems involve the stochastic generation, manipulation and removal of individual particles from a global system. Particles in the system can also be affected by external forces, leading to the simulation of stochastic behaviour. Fluid simulations are based around the idea of iteratively solving density and velocity grids, where

every cell in the grid has a density and velocity value dependent on the neighboring cells' values. Although they are fundamentally different techniques, fluid simulations and particle systems are often used to model similar natural phenomena. Thus they will henceforth be collectively referred to as stochastic motion simulators.

These techniques have been shown to be very useful in audio-visual performance and installation contexts. Systems and code-libraries for both fluid simulations and particle systems usually present global parameters to the user that can be altered in real-time to vary the properties of the system over time, which makes them ideally suited to real-time parameter mapping. For example the fluid simulation system developed by Forbes, Höllerer, and Legrady (2013) was intentionally designed for use in multimedia arts and performance settings. It exposes multiple parameters that control the quality of the fluid. These include *momentum*, *directionality* and *angularity*, which control the directional movement of the fluid through the space; *viscosity*, which controls the rate at which energy is removed from the system, and *sensitivity*, which controls how the fluid reacts to user input.

In the context of audio-visual performances, a common mode of interaction with stochastic motion simulators is gestural interaction. Renaud et al. (2013) developed the '3dinmotion' system, which uses motion tracking data of multiple users to drive sonification and visualisation algorithms in parallel (Renaud, Charbonnier, and Chagué 2014). The visualisation algorithm uses multiple particle emitters to visualise motion data.

This type of congruent mapping to both aural and visual parameters from gesture and motion parameters is a commonly explored area in the context of stochastic motion simulators for audio-visual performance (Momeni and Henry 2006). Johnston (2013) developed a system that tracks the motion of performers in real time and uses their movements to influence a fluid simulation algorithm. The algorithm is then used both to render visuals and to generate audio. The audio is generated using concatenative synthesis. Macro-cells containing multiple individual cells from the fluid velocity grid are used to control the individual grains in the concatenative synthesis system. Average velocity of macro-cells is used to control the gain of individual grains, which continuously play very short looped samples. Grains are organised by timbre such that neighbouring cells in the grid control grains with similar sound characteristics.

Hsu (2011) describes the use of a particle system in the context of congruent gesture-to-audio and gesture-to-visual mappings. User input data such as motion and placement of attractors controls the particle system's behaviour. The user's gestural input is simultaneously used to generate audio. Descriptive parameters are exposed from the audio generation algorithm (e.g. 'brightness', 'loudness', 'duration') and these can be controlled using the gestural input.

Although stochastic motion simulators in real-time performances are most commonly used in congruent audio-visual synthesis, they have also been used to directly visualise audio features.

Fonteles, Rodrigues, and Basso (2013) developed a system for music visualisation that uses properties of particle systems to visualise different musical qualities. Pitch is mapped to particle colour and volume to particle size. Differences in timbre are represented only at the level of instrument category, and each individual instrument produced emissions from individual particle emitters.

Hsu (2011) also describe the use of his particle system in a real-time audio visualisation context, where the particle system reacts to real-time audio descriptors. The audio descriptors used are those

from the Zsa.Descriptors library (Malt and Jourdan 2008). Limited detail is given on exactly how these descriptors are mapped, but emphasis is given to multidimensional mappings that produce visualisations that can be 'open to interpretation'. One example given of such a mapping is a 'slow ... loud sonic gesture' producing a 'large tidal current' in the particle system.

2.6 Conclusion And Summary

The timbre of a sound is a difficult quality to define. The standard definition separates it entirely from pitch and loudness - an assumption that is being called into question by ongoing research. The advent of digital synthesis of sound has highlighted the continuous nature of timbre (as opposed to the concept of static timbres, linked to individual sources), and has motivated the study of timbre in isolation, separate from the notion of 'sound source'.

Electroacoustic music is a type of sound art that gives prominent focus to timbre and timbre variation as compositional structure. The concept of spectromorphology is an attempt to formalise and categorise the various compositional concepts used in electroacoustic works. Spectromorphology is concerned with spectral properties, structures, and events. In this way, electroacoustic music and spectromorphology provide a new way of defining timbre. It can be understood as all of the spectral content of a signal and its variation over time.

A lot of existing research into timbre perception has been conducted in the context of constant pitch, volume and duration. However, multiple studies have suggested that the perception of pitch and timbre can interfere under certain circumstances. For example, without an extended tonal context, timbre variation seems to affect pitch perception. Similarly, in the presence of varying pitch, contextual timbre perception can be diminished. There is also evidence suggesting that such interference can be dependent on level of musical training or experience.

The technique of multidimensional scaling (MDS) can be used to study how sound is perceived in the context of identical pitch, loudness and duration. Timbre spaces are the result of dimensionality reduction on similarity ratings of sounds, using MDS. Timbre spaces can model how different contexts, listener traits, and specific sound characteristics influence the perception of timbre. They can also be used as a model for the design of novel control interfaces for audio manipulation, where manipulation of timbre is achieved through direct exploration of a physical space. One of the main issues in designing such interfaces is the representation of timbre spaces with over 3 dimensions.

When describing sounds that have no obvious source-cause, listeners will often turn to metaphor. There is a large body of evidence suggesting that the metaphorical language used by listeners is mostly visual and tactile, with descriptors being material, textural and physical in nature. The techniques of semantic differential and verbal attribute magnitude estimation (VAME) are used, combined with dimensionality reduction techniques such as principal components analysis (PCA), in order to extract the most salient semantic descriptors of timbre. One of the biggest issues involved in such studies is the choice of semantic descriptors that is initially presented to participants. By identifying correlations between acoustic features and semantic descriptors, a novel visualisation paradigm for musical timbre could be developed. This would alleviate the issues involved with the representation of high dimensional information, when visualising timbre spaces.

There seem to be numerous acoustic features that account for differences in timbre and have

an effect on our perception of timbre. Efforts to quantify these acoustic properties of sound have come from various different fields including MIR, musicology and psychology. An extensive list has been provided by Peeters et al. (2011). This list is mainly broken up into temporal and perceptual features. Research into auditory perception has identified the 'modulation spectrum' as a rich source of spectro-temporal information within a signal. The various dimensions of timbre spaces can be compared to specific acoustic features using regression analysis to identify correlations between features and dimensions. Regression analysis can also be performed on semantic descriptors, in order to identify correlations between acoustic features and semantic descriptors. By combining timbre spaces from MDS studies, salient semantic descriptors for timbre resulting from semantic differential studies, and correlated acoustic features, a perceptually motivated mapping strategy could be developed that links specific acoustic features to their visual metaphors (given that many of the semantic descriptors are visual, textural and material in nature). This could provide the groundwork for a new perceptually motivated approach to timbre representation. Such representation techniques could be used to improve interface design for sound manipulation interfaces.

The identification of quantifiable acoustic timbre features combined with the use of timbre variation as compositional structure suggest that an improved definition of timbre would refer to individual audio characteristics directly, rather than collectively as 'everything that isn't pitch or loudness'. Similarly, the identification of specific acoustic timbre features suggest that timbre could be accurately represented by charting the value of these features over time. The level of accurateness here is determined by the number and variation of different audio features that are used. Representation techniques ultimately have to weigh-up the level of detail vs. the usability and context. Attempts have been made to develop such timbre representation methods based on the visualisation of individual audio features. One of the key issues in the development of such representation techniques is the mapping of acoustic features to visual features. The use of arbitrary mappings can hinder the generalisability of representation techniques. Similarly, generalisability is limited by the fact that different representation methods make use of different mappings. In the context of electroacoustic music, numerous attempts have been made to develop timbre notation systems. These systems are usually subjective, abstract and context dependent. They represent timbre in perceptual terms, illustrating the perceptual impressions that result in a human listener from the presence of specific spectral properties and events. The use of perceptually motivated mappings from acoustic timbre features to visual properties could provide the basis for a perceptually motivated timbre representation paradigm. This could be used to link the low-level descriptiveness of data-driven representation methods with the illustrative perceptual qualities of spectromorphological timbre notation methods.

Timbre notation techniques and data-driven temporal timbre representation methods are based around charting timbre variation visually along some form of timeline. Timbre can also be represented in real time, as a data-driven animation that is either generated in parallel with – or in response to – real-time timbre variation. Numerous examples of audio-visual performances and installations show how timbral qualities of the audio can be used to influence – or directly drive – the visual components. The development of a perceptually influenced data-driven timbre representation paradigm could be used in the context of live performances to draw attention to the use of timbre variation and as a compositional and performative device.

Chapter 3

Methodology & Experimentation

The previous chapter gave a detailed account of existing research into the definition, perception, measurement and description of timbre. Given this technical background, this chapter describes the experimental methods used to further investigate the description and representation of timbre. These experiments have been carried out in order to approach the key motivating issues and research questions being addressed in the thesis, as described in section 1.3.

The experiments outlined in this chapter set out to develop mapping strategies between acoustic timbre features and semantic timbre descriptors. Such mapping strategies are central to the key problems being addressed in this thesis. The first problem concerns the perceptually motivated representation of timbre and timbre variation. The identification of robust, perceptually salient mappings between acoustic features and semantic descriptors facilitates the use of visual properties to represent timbral qualities. Variation in visual properties, either temporally (for real-time contexts) or along some visual dimension (in other contexts) allows temporal variation in timbral qualities to be represented. The second problem concerns interface design. Again, robust mappings between acoustic features and semantic descriptors provide the basis on which more intuitive visual tools can be built. The use of visual qualities to represent timbral qualities introduces the possibility that these visual qualities can be physically manipulated in order to alter the timbral qualities. This provides a perceptually-oriented approach to interaction as opposed to individual parameter manipulation through interaction with technically labeled control elements.

The first section in this chapter describes a preliminary user study that was carried out in order to investigate the visualisation of timbre features using 3D animated visual stimuli. The experimental setup is described and results are reported and discussed. The graphical representation methods explored in the preliminary study form the basis of a real-time timbre representation system described in a later chapter. The second section describes the development of a mapping tool that can be used to map synthesis parameters to visual parameters. This is described along with key implementation details. The tool is presented as a specific approach to abstract timbre representation in digital synthesis tools and is used to introduce a specific form of representation referred to as 'synthesis representation'. A discussion is given on the differences between visualising synthesis parameters and visualising audio features. The final section describes an online interactive user survey that was implemented in order to investigate participant preferences for

different visual timbre mappings in the context of coloured amplitude waveform visualisations. The algorithm behind the coloured waveform technique is outlined. This is followed by a description of the various visual properties that are used in the coloured waveforms. The online survey is then described with reference to the implementation and results, and is discussed in terms of implications for visual timbre mappings. This timbre visualisation technique forms the basis of a novel timbre representation system discussed in a later chapter.

3.1 Animating Timbre: A Preliminary User Study

A preliminary user study was conducted in order to investigate user preferences for different visual mappings of acoustic timbre features.

As mentioned previously in section 2.4.2, a key issue with many existing user studies into preferred visual mappings for acoustic timbre features is that they have used static, 2D images as visual stimuli. Therefore, as well as investigating user preferences for visual mappings, one of the main objectives was to expand on previous studies by making use of complex 3D animated objects as visual stimuli.

The aural and visual stimuli, discussed later in this section, were influenced by previous studies investigating timbre perception and description. Specifically, the aural descriptors used were taken directly from Caclin et al. (2005) and the visual descriptors are based on those identified by Zacharakis, Pastiadis, Papadelis, et al. (2011).

3.1.1 Participants

18 participants took part in the study (mean age = 28.8, 9 female). 11 had received at least some formal musical training, and 6 were regular users of audio production/synthesis software.

3.1.2 Stimuli

Auditory Stimuli

Audio tones were generated by additive synthesis using Supercollider. The fundamental frequency was kept constant for each tone, at 311 Hz (E♭4). The audio parameters used in the study, along with their values, were based on those reported by Caclin et al. in their study on acoustic correlates of timbre space dimensions (Caclin et al. 2005). In their study, Caclin et al. identified 3 salient acoustic timbre descriptors: *attack time*, *spectral centre of gravity* (SCG) and *even harmonics attenuation* (EHA). The same 3 features were used in this study. The attack time varied logarithmically between 15ms and 200ms, as it has been suggested that listeners use a logarithmic scale when using attack time to discriminate between timbres (McAdams, Winsberg, et al. 1995). Caclin et al. provide methodologies for varying the SCG and EHA. The same methods were used

in this study. SCG was manipulated using

$$A_n = k * 1/n^{\alpha} \tag{3.1}$$

where A_n = the amplitude of the n^{th} harmonic. The value of α determines the value of the instantaneous spectral centre of gravity. SCG varied linearly between 1400 Hz (4.5 in harmonic rank units) and 664 Hz (2.1 in harmonic rank units). This was achieved by varying α between 1.23 and 2.07. EHA was controlled by changing the level of the even harmonics relative to the odd harmonics using

$$Eh_n = Oh_n * 10^{\beta/20} \tag{3.2}$$

where Eh_n = the amplitude of the n^{th} even harmonic and Oh_n = the amplitude of the n^{th} odd harmonic, and β = the relative change in volume (in dB). During experimentation, β ranged from -8 to 0.

Visual Stimuli

In their investigation into semantic descriptors of timbre, Zacharakis et al. observe that it seems reasonable to identify musical timbre using verbal associations to physical objects' properties (Zacharakis, Pastiadis, Papadelis, et al. 2011). As mentioned previously in section , existing research has identified properties of texture and shape as salient visual correlates of timbre features. Colour and position have been identified mainly as correlates of pitch and loudness. For this reason, no colour was used in the animations and the position remained constant.

Each animation consisted of one 3D rendered polyhedron. The polyhedron was modelled using geodesic subdivision, with an icosahedron as the seed shape (Lounsbery, DeRose, and Warren 1997). The subdivision depth (spherical resolution) was one of the parameters controlled during animations, and this ranged from 0 to 6. The polyhedron was modelled within a unit sphere, and triangular pyramid 'spikes' protruded from each surface face. The length of the spikes was controlled during animations, and ranged from 0 to 1. The two other visual parameters that were varied during animations were brightness and opacity. The visual parameters used in this study (spherical resolution, spike length, brightness and opacity) were based on the 3 factors identified by Zacharakis, Pastiadis, Papadelis, et al. (2011), namely *volume and wealth* (spherical resolution), *brightness and density* (brightness/opacity), and *texture and temperature* (Zacharakis, Pastiadis, Papadelis, et al. 2011). Various surface textures were possible on the polyhedron through a combination of different spherical resolution and spike length values, as demonstrated in figure 3.1. The animations were implemented using C++ and OpenGL.

3.1.3 Experimental Procedure

Participants were asked to complete two separate tasks, both of which involved giving indications of their preference for different audio-visual mapping strategies. For each task, participants sat in front of a 15 inch laptop display and were equipped with headphones. Participants used simple graphical interface panels (developed in Supercollider) on the right of the screen in order to listen to different audio tones and cycle though different mapping strategies. The resulting visualisations were displayed in a large window on the left of the screen.

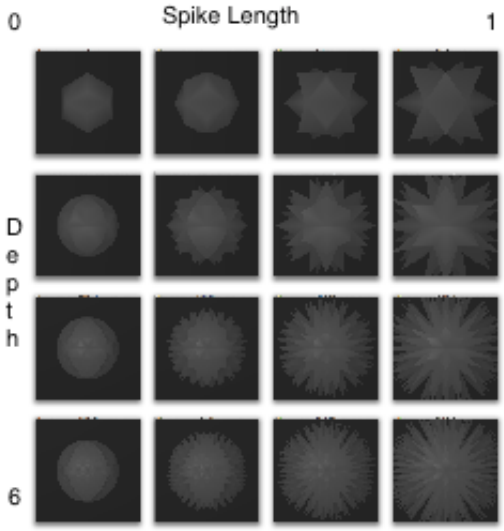


Figure 3.1: Rendered polyhedra with varying spherical resolution and spike length values.

Task 1: Individual Parameter Mapping

Objectives

Task 1 was designed in order to introduce participants to the different audio and visual parameters, and to record their favourite to least favourite visual mappings for each audio parameter.

Procedure

During task 1, a single audio parameter changed while the others remained constant. For each audio parameter, participants were presented with three consecutive tones. The selected audio parameter was increased between each tone. Parameter values increased between the values reported previously in section 3.1.2. Each tone produced a resulting visualisation in which one of the visual features changed along with the audio feature, according to the selected mapping. Participants could observe an audio-visual stimulation by pressing the 'play' button in the control interface. There were also buttons to change which audio feature was being varied, and which visual feature the audio feature was mapped to.

For each audio feature, participants were asked to cycle through the different visual mappings and rank them from 3 (favourite) to 0 (least favourite). Thus, they constructed a *preference table* for the various mapping possibilities. An example of a participant's preference table is given in table 3.1. The participants filled in their preference table as they progressed through the task. They were able to observe stimuli and update their preference ratings as many times as they needed.

	Res	Spike	Bright	Opacity
Attack	3	2	0	1
SCG	1	3	2	0
EHA	1	2	3	0

Table 3.1: An example preference table for a participant (3 = favourite, 0 = least favourite).

Results

The 'Borda count' can be used to analyse the results of a preference vote. It is basically a weighted count of votes, weighted on the preference of each vote. In this case, for each audio parameter, every visual mapping is given a Borda count. Every time a participant gives a preference rating to a visual mapping, the value of that preference rating is added to the overall Borda count of the visual mapping. A higher Borda count therefore implies a greater preference.

Figure 3.2 shows the Borda scores for all visual mappings, for attack (blue), SCG (red) and EHA (green). Figure 3.3 shows the relative preferences of the mappings, according to their Borda

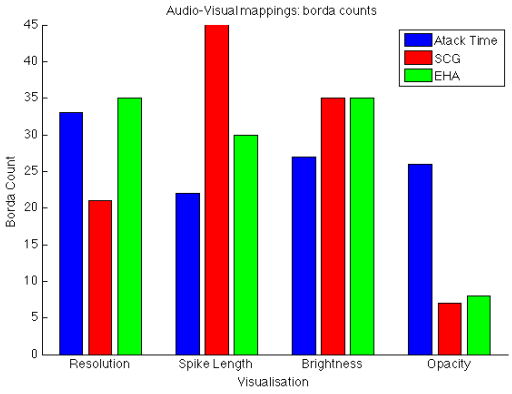


Figure 3.2: Borda counts for each visual mapping, for all audio features.

counts. The point size represents the overall popularity of each audio-to-visual mapping, compared to the other options.

Another way to analyse the results of a preference vote is to use a Condorcet method. This pits every candidate (e.g. mapping) against one another in pair-wise hypothetical contests. If one candidate wins every contest, they are considered the Condorcet winner. In this case, a certain visual mapping is the Condorcet winner for an audio feature if it has a higher (or equal) Borda count than all other possible visual mappings for that audio feature.

Each audio feature had a visual mapping that emerged as the Condorcet winner when all visual mappings were compared using a Condorcet method, as shown in table 3.2.

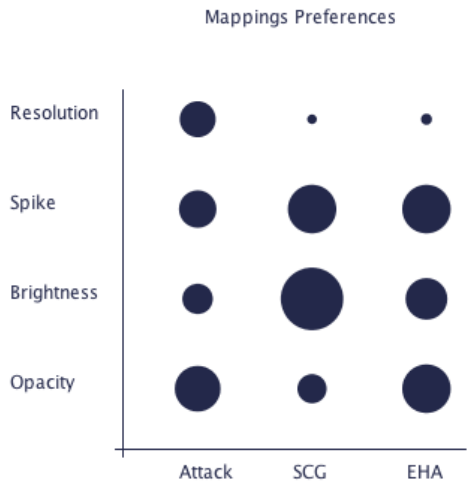


Figure 3.3: Borda counts of each audio-visual mapping. Point size = relative popularity of an audio-to-visual mapping in comparison to the other options.

	Visual Mapping
Attack	Resolution
SCG	Spike Length
EHA	Brightness

Table 3.2: Condorcet winner visual mappings for each audio feature.

Task 2: Multiple Parameter Mapping

Objectives

Task 2 involved all of the audio and visual mappings changing at once. The task was designed to encourage the participants to explore different global mapping strategies and to record their recommended optimal global mapping strategy. A key objective was to ascertain to what extent mapping preferences change, if at all, when multiple parameters are varied together.

Procedure

During task 2, participants listened to short audio tones in which each of the audio features were randomised. The SCG and the EHA were also varied *during* the audio tones, using randomised linear envelopes. Such audio tones produced visual animations where the visual features of the polyhedron varied smoothly and in direct response to the audio, according to the mapping configuration. Participants could observe a randomised audio-visual stimulus at any point, by pressing 'play' in the control interface. There were also four buttons for each visual parameter, which allowed the participants to change which audio feature was controlling that visual parameter.

Using the preference tables from task 1 (table 3.1), suggested optimal mapping strategies were constructed. Table 3.3 gives an example. These were used to construct the initial mapping strategies in task 2. Participants were then encouraged to explore different mapping configurations and evaluate them by observing some randomised audio-visual stimuli. Ultimately participants indicated what they thought was the *optimal* mapping configuration by filling in an optimal mapping configuration table (see table 3.4 for an example). The mapping configuration consisted of each visual feature being controlled by a single audio parameter and multiple visual features could be controlled by the same audio feature.

	Suggested Visual Mapping
Resolution	Attack
Spike	SCG
Brightness	EHA
Opacity	Attack

Table 3.3: An example suggested optimal mapping strategy for a participant (using the results from task 1).

	Optimal Visual Mapping
Resolution	EHA
Spike	SCG
Brightness	EHA
Opacity	SCG

Table 3.4: An example optimal mapping strategy for a participant (from task 2).

Results

Participants' *suggested* mapping configurations from task 1 were compared to their *'optimal'* mapping strategy from task 2. Comparing the two tables, a measure of the difference between the preference table and the optimal strategy can be evaluated. This difference is calculated as the total number of mappings in the 'optimal' strategy that differ from the suggested mappings from the preference table. For example, the difference between the suggested mapping strategy in table 3.3 and the optimal mapping strategy in table 3.4 would be 2. This value gives an indication as to what extent a participant's preferences from task 1 varied, after exploring global strategies during task 2.

In total, 14 (78%) of the participants' optimal strategies differed from their suggested strategies. 8 of these changed by 1 mapping, 2 changed by 2 mappings, and 4 changed by 3 mappings. Figure 3.4 shows how often each visual attribute changed, between suggested and optimal mapping strategies.

In total there were 12 unique 'optimal' mapping strategies that emerged from the study, the most popular being common to only 3 participants (shown in table 3.5). 4 optimal strategies were common to 2 participants, and the other 7 were unique to individual participants.

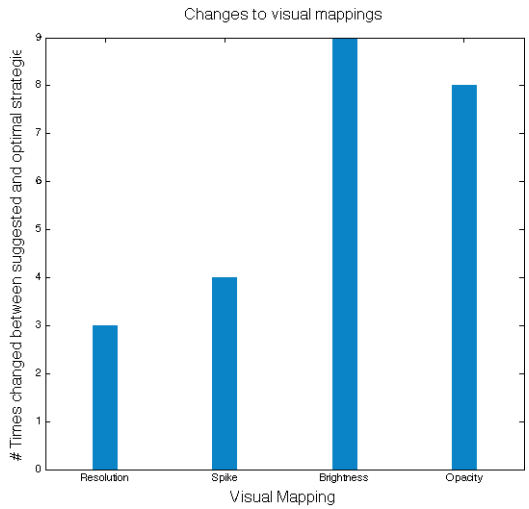


Figure 3.4: Number of times each visual mapping changed between suggested and optimal mapping strategies.

	Optimal Visual Mapping
Resolution	EHA
Spike	SCG
Brightness	SCG
Opacity	Attack

Table 3.5: 'Most popular' optimal mapping strategy (common to 3 participants).

3.1.4 Discussion

Task 1 identified preferred visual mappings for attack time, SCG and EHA. The existence of a Condorcet winner for each suggests that there was general agreement on which visual features best represented which audio features in isolation, namely attack to spherical resolution, SCG to spike length, and EHA to brightness. These results are contradictory to many previous studies that have found SCG to be highly correlated with the 'brightness' of a sound. However, figure 3.3 identifies a very wide spread between preferred mappings, especially for attack time and EHA. Attack time, SCG and EHA have been reported as salient axes for timbre discrimination (Caclin et al. 2005). Volume, texture and brightness have been reported as salient verbal descriptors of timbre variance by Zacharakis, Pastiadis, Papadelis, et al. (2011). The results from task 1 can possibly indicate which verbal descriptors relate to which acoustic features (where spherical resolution relates to volume and spike length relates to large-scale texture). However, larger studies are required to confirm or refute these findings. It is also possible that there are n-m visual-to-auditory correspondences rather than 1-1. Some of the participants commented that the mapping from SCG and EHA to opacity should have been inverted. This may have had an effect on the popularity of opacity as a visual mapping.

Task 2 identified variation between many individual participants' isolated mapping preferences and their global optimal mapping strategy. This suggests that mapping preferences change when

multiple parameters are in flux. Again, larger studies are required in order to further evaluate this suggestion. The variance in individual preferences could be due to the psychoacoustic perception of timbre, as it has been suggested that the salience of acoustic timbre features depends on the number of features in flux (Caclin et al. 2005). It is interesting to examine the total number of times (across all participants) each individual visual mapping was changed between suggested and optimal mapping strategies. Figure 3.4 indicates that there was more disagreement on the use of brightness and opacity as visual mappings than there was for resolution and spike length.

Task 2 was designed to encourage exploration in the participants, such that their preferences were their own, rather than one of a limited number of options presented to them. To facilitate this, the number of possible mapping configurations was left intentionally large. However, this resulted in a large cognitive load for the participants. Thus, despite the objective of avoiding 'right or wrong answers,' it is possible the large cognitive load resulted in the task feeling 'too difficult' for some participants. The measure of 'difference' between mapping strategies, as defined in section 3.1.3, can only be used as a very general indication of difference, since the differences being measured are perceptual and their magnitudes vary.

There was a large degree of variance between the different participants' suggested global optimal mapping strategies. This supports the idea that mapping preferences change as the number of mappings increases, and suggests ambiguity in preferred visual mapping from listener to listener.

The audio features used in this study were inspired by those used in a previous study (Caclin et al. 2005), however, they only constitute a limited sub-set of timbre features.

3.1.5 Conclusion

The aim of this study was to combine findings about verbal timbre descriptors and acoustic timbre features and explore preferred mappings between the two. In the case of isolated mappings there seem to be clear audio-to-visual mapping preferences. When multiple mappings are considered, no clear preference emerges, and preferences sometimes change from the isolated case.

The other key aim of the study was to make use of 3D animated visual stimuli – something lacking from other similar user studies. The audio-visual stimuli were based on the deformation of a spherical mesh in order to produce varied visual shapes and textures in response to changes in timbre features. This technique has not been used in other such studies, and provides the basis for the development of a novel interaction tool described later in the thesis.

The findings of the user study suggest that any graphical applications exploiting perceived associations between auditory timbre and visual form may benefit from customisable mapping strategies. However, the participants involved in this study were not widely representative of prospective users of such systems (with only 6 being regular users of audio production software). In researching systems intended for use in digital sound design environments, it would be beneficial to use larger, more homogeneous participant groups with more experience of said environments.

This study used only monophonic, synthesised audio tones. It would be beneficial to also study participant preferences in the context of natural/acoustic and/or polyphonic audio stimuli. Another potential improvement for future studies would be to test the reliability of preferences. This would involve presenting the same stimuli at various times and checking whether preferences

remain consistent.

Larger studies are required to confirm or refute the findings reported here. Studies with larger participant numbers could also help identify whether there are different categories of preference (e.g. whether certain mapping combinations usually go together). The acoustic and visual features used in this study were based on findings reported elsewhere, but future studies may benefit from using larger parameter sets.

The main drawbacks with this study are similar to most other existing studies of this type, as mentioned in section 2.4.2, namely, the small participant pool and the controlled lab environment. In order to provide more generalisable and reliable results, future studies should use a larger number of participants and should be implemented in varied settings.

3.2 Synthesis Parameter Representation

The study described in the previous section used additive synthesis such that key timbre features were directly controllable and could be directly mapped to visual parameters. In a tool such as Equator, the synthesis methods are more complicated, and do not necessarily provide direct access to timbre features. For example the filter cutoff has differing effects on the timbre depending on the filter type that is being used. Visualisations using synthesis parameters in a tool like Equator therefore require the ability to create complex n-m mappings from synthesis parameter to visual parameter. The results of the user study described in the previous section showed variation in individual user preferences depending on the number of parameters in flux. This suggests that customisable mappings may be appropriate in timbre representation tools. This section describes a tool that was implemented in order to facilitate rapid prototyping of complex visualisation strategies for synthesis parameters in Equator. The tool was also implemented to test the viability of customisable mappings.

The key objective here was to research and develop a new form of representation for the Equator interface that makes use of textural, material and physical properties in order to represent the engine's configuration in a visual, semantically meaningful way. This is related to the research problem of interface design in digital tools for timbre creation and manipulation.

An early approach to this issue involved the development of a generative visualisation tool by a colleague at ROLI. A corresponding mapping tool was developed that links the visualisation tool to the Equator engine. This section gives a brief overview of the generative visualisation tool before describing the mapping tool in detail. A discussion is given on the use of the mapping tool and the generative visualisation tool to create visual representations of Equator presets. This discussion will introduce the concept of 'synthesis visualisation'. This section concludes with a discussion on the merits and drawbacks of using synthesis parameters (rather than audio features) to drive visualisation.

3.2.1 'Soundflake' Tool

The Soundflake generator tool was developed by Felix Faire, a colleague at ROLI. It enables the creation of widely varying graphical visualisations with distinct textural and material properties. It is based on a fragment shader that colours pixels depending on their position and a number of control parameters. The fragment shader produces radial forms with many textural and colour-based distortions and effects. These visual effects are controlled by a number of control parameters. For example, the 'blur' parameter applies a blurring effect to the radial image. The 'sparkle' parameter can be used to control a shimmering effect. The 'twist' parameter controls the extent to which radial texture effects are twisted. The different visual forms produced by the tool are referred to as Soundflakes due to their wide variability and their radial nature. Some examples of Soundflake images are shown in figure 3.5.

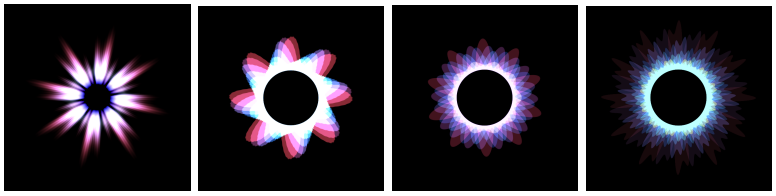


Figure 3.5: Examples of Soundflake images.

3.2.2 The 'EMapper' Tool

The Equator Parameter Mapper (henceforth 'EMapper') tool was developed in order to facilitate rapid prototyping of different mapping strategies from Equator synthesis parameters to control parameters of the Soundflake generator. The aim was to facilitate a way in which the configuration of the Equator synthesis engine could be represented visually as a unified graphical form, rather than as a list of technical engineering parameters and associated values. Figure 3.6 shows the EMapper panel laid over the Equator interface.

Mappings

The EMapper tool was developed as a panel within the Equator interface. The EMapper panel presents the entire list of parameters included in the Equator synthesis engine in a column on the left. In a similar column on the right the visual control parameters of the Soundflake generator are presented. Connections can be made between the parameters in order to construct a mapping strategy from synthesis parameters to visual parameters. By clicking on a synthesis parameter followed by a visual parameter, a direct mapping is created. By clicking on a visual parameter first and then on a synthesis parameter, an inverse mapping is created such that high synthesis parameter values produce low visual parameter values. When multiple synthesis parameters are linked to the same visual parameter, the average of the synthesis parameters is taken and mapped to the visual parameter. Mapping strategies can be saved and recalled. Binary files with the extension '.emap' are used to store the mapping strategy configuration. As shown in Figure 3.7,

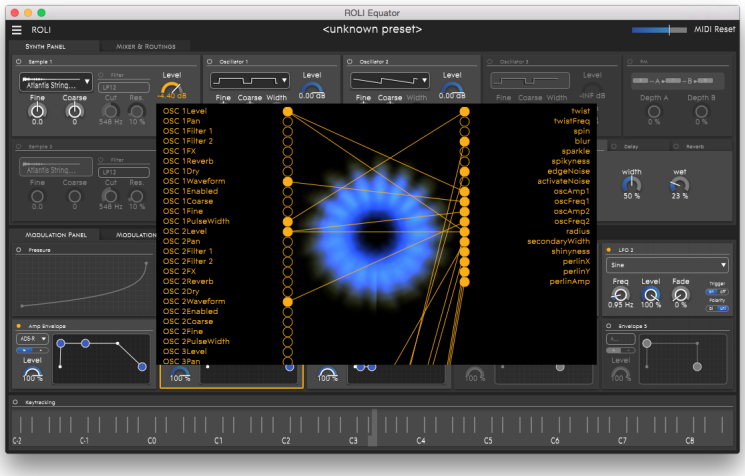


Figure 3.6: The EMapper panel within Equator.

there are four buttons at the bottom of the panel for showing/hiding the mapping configuration, updating the graphics shaders in real time, and saving and loading mapping configurations.

Synthesis Parameter Tracking

There are two ways in which the synthesis parameters that the Equator sound engine provides can be accessed while the system is running. The parameters have base values that are set using the sliders in the interface. Accessing these values provides the configuration of the sound engine 'at rest' without any modulation applied. There is also the possibility of tracking the values after all modulations have been applied. This provides access to the configuration of the sound engine 'in flux' as modulation is affecting the parameters in real time. These two types of parameter tracking can be used in different contexts, as outlined later during the discussion.

Run-Time Shader Compilation

As well as the construction of mapping strategies, the EMapper tool allows live updating of the fragment shader that implements the Soundflake generation, such that the visual algorithm can be tweaked and updated in real time. The 'update shaders' button in the EMapper panel re-compiles the fragment shader and updates the graphical output. This proved very useful in varying the global scale and thresholds of certain visual effects, for example.

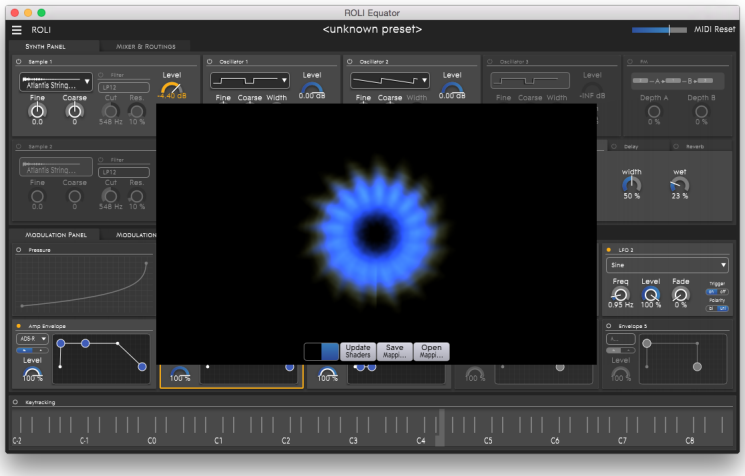


Figure 3.7: EMapper buttons panel.

3.2.3 Discussion

The EMapper tool facilitates the development of mapping strategies between synthesis parameters and visual effects. These mapping strategies are based on the use of semantic descriptors and visual, textural properties to describe the configuration of the underlying sound engine and to characterise the kind of sound output that the configuration produces.

Complex Mappings

As mentioned previously in section 2.4.2, multiple studies have found semantic descriptors of timbre to be correlated with multiple dimensions in a perceptual timbre space. It is therefore important for tools like EMapper to facilitate complex n-m mappings from synthesis parameter to visual parameter. EMapper facilitates these kind of mappings by taking the average value of all synthesis parameters and applying this to the destination visual parameter. One way in which the EMapper tool could be improved and extended is to allow for more complex rules in the way synthesis parameters are combined: for example additive and non-linear combinations.

Synthesis Visualisation

The Soundflake generator has a specific algorithm that is used to generate varied visual output. This visualisation algorithm can be considered as a 'visuals engine' that complements the Equator sound engine, by providing an animated visual analogy to the generated sounds.

The Equator engine is capable of producing a wide array of different timbres, and each configuration of the various synthesis parameters denotes a single point in the high dimensional timbre

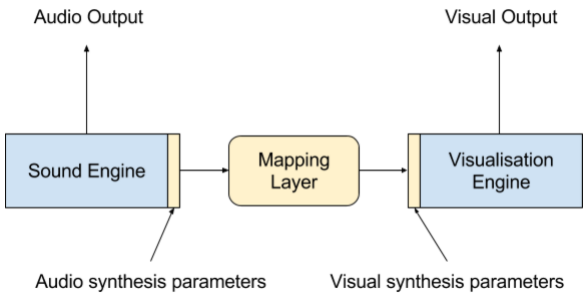


Figure 3.8: Synthesis visualisation. Congruent sound and visual synthesis, driven by a parametric mapping layer.

space afforded by the engine. The modulation of synthesis parameters (e.g. by touch parameters or modulators like LFOs) then facilitates the movement to neighbouring regions in the timbre space. Thus, Equator presets (configurations of synthesis parameters combined with mappings from modulators to synthesis parameters) constitute regions or 'sub-spaces' within the overall timbre space.

In a similar way, the Soundflake generator tool is capable of producing a wide array of different visual effects and each configuration of the visuals engine produces a unique graphical visualisation. The static visualisations thus serve as representations of particular presets, or points in the timbre space of the Equator synth. These representations are produced by tracking the sound engine configuration 'at rest' (as discussed previously in "Synthesis Parameter Tracking").

By tracking the synthesis parameters 'in flux' after modulation has been applied, mapping strategies can be set up such that the modulation of Equator synthesis parameters produces procedural animations in the visualisation that visually represent the sub-space afforded by the Equator preset. These can be thought of as 'icons' (either static or animated) that indicate a specific preset configuration.

In a similar way, by tracking the synthesis parameters in flux during a performance (while notes are being played and user data is modulating the parameters), the mappings can produce real-time reactive animations that provide a visual accompaniment to the performance.

For the purposes of discussion later in the thesis, this form of parameter-driven procedural visualisation is labeled as 'synthesis visualisation'. The term comes from the fact that the mappings make use of synthesis parameters rather than direct audio features. The visualisation algorithm is based around a similar structure to the sound synthesis algorithm and there is a mapping layer between them that links the two routines, as shown in Figure 3.8. Synthesis visualisation is effective for representing complex parameter spaces and underlying timbre spaces visually. Presets can be effectively distinguished since the different configurations of the sound engine are directly linked with different forms in the resulting visualisation. Procedural animations can be simply generated that represent the particular region of the timbre space that a preset covers. Animations can also accompany a performance in real time in order to visualise the performative exploration of the timbre space.

M. Blackburn (2011) discusses a framework for electroacoustic composition based around visual sound shapes that represent various spectral events and characteristics. In a similar way, these visual synthesis representations could be used to develop a form of visual timbre composition. This kind of composition would be perceptually (rather than parametrically) driven. M. Blackburn (2011) talks of 'visual sound sculpting' through the selection and manipulation of individual sound units. The visual synthesis representations could be used in a similar way. They can be used as individual synthesis icons or animations in order to indicate the timbre of particular presets. These presets can then be combined and played together, producing combined visual effects. Since the visual properties represent synthesis parameters, an immediate visual synthesis language emerges such that the manipulation of synthesis parameters produces direct visual representations of how this affects the timbre.

3.2.4 Conclusion

This section has described the development of a mapping tool whereby synthesis parameters can be mapped to visual parameters to create semantically relevant visual effects that illustrate the resulting timbral characteristics of a given synthesis engine configuration (preset). As an extension to the Equator interface, the EMapper tool facilitates rapid prototyping of mapping strategies for Equator presets. The main objective behind the development of the EMapper tool was to facilitate the semantic representation of the Equator sound engine's configuration, using textural and visual qualities rather than technical labels. In order to achieve this, the EMapper tool allows direct and inverse n-m mappings between audio synthesis parameters of the sound engine and visual synthesis parameters of a visual synthesis engine. This process is referred to as synthesis visualisation. Synthesis visualisation is an abstract form of timbre representation, since neither the aural nor the visual parameters describe the actual audio or image data but rather control the way in which the data is produced or altered. The use of synthesis parameters rather than real-time audio features lowers CPU load since there is no need for intensive real-time audio processing.

Although the initial investigation into parametric representation has been promising, the use of synthesis parameters is inherently specific to particular synthesis methods, techniques and tools. Similarly, the visualisations are entirely dependent on the visualisation tool that is used. This thesis has discussed the use of the Soundflake generator but any type of procedural visualisation algorithm could be used. This means there is a lot of room for experimentation and development with the synthesis visualisation approach, but also means that any system that is developed will be very context dependent. The context-specific nature of the method was inappropriate for the development of more general techniques outside the context of Equator. The rest of the techniques and systems described in the thesis therefore focus on low-level audio features extracted from the audio signal through signal processing methods.

3.3 Temporal Timbre Representation

The tools described in the previous section approach the representation of timbre in an abstract way, by visualising synthesis parameters using a bespoke visualisation algorithm intentionally designed to produce semantically relevant visual effects. This section returns to the use of direct low-level audio features in visual timbre representation. Rather than using the features directly

as synthesis parameters (as in the preliminary user study described earlier in the chapter), the work described in this section is based on the real-time extraction of timbre features through audio signal processing techniques. The tool described in the previous section was designed in order to tackle the issue of interface design in digital tools. It provides a method by which the configuration of a synthesis engine can be represented semantically using visual properties. This section describes the development and evaluation of a more general technique for audio representation. This technique was developed in order to demonstrate the use of perceptually motivated mappings for timbre representation.

One of the most common visualisation tools used in audio production environments is that of the amplitude waveform. Gohlke et al. (2010) provide a list of tasks for which amplitude waveform visualisations are insufficient. In the context of tools for sound design – where sounds are created from the ground up – the most important drawback of the standard amplitude waveform is its inability to represent timbre features beyond those related to the amplitude envelope. One way in which this drawback can be alleviated is to use something like a spectrogram representation, which shows the temporal evolution of spectral information over time. However, the use of spectrograms limits the ability to read amplitude information over time. The novel visualisation technique presented in this section is known as the 'Sound Signature' technique. Sound Signature images visualise the spectral evolution of sounds using colour and texture.

An experiment will be discussed in this section that was conducted in order to evaluate participant preferences for different timbre visualisation strategies in the context of Sound Signature visualisations. One of the key objectives of the study was to identify the presence or absence of preference for Sound Signature representations over standard amplitude waveform visualisations. The secondary objective was to identify general preferences for specific visual mappings in the context of Sound Signature visualisations.

This section is concerned primarily with experimentation that was conducted using the Sound Signature technique. The development of an integrated extension to the Equator interface using this technique is described later in the thesis. In this section, the algorithm behind the technique is described. This is followed by a description and discussion of an experiment that was conducted using Sound Signature visualisations in order to evaluate various visual mappings of timbre features.

3.3.1 Algorithm Description

The algorithm behind the Sound Signature feature can be broken into three key components: windowing and down-sampling of audio data, timbre feature extraction, and visual rendering.

Windowing & Down-sampling

The Sound Signature technique is similar in to that of TimbreGrams (Tzanetakis and P. R. Cook 2000), Comparisonics coloured waveform displays (Rice 2005), and the BStD method (Malt and Jourdan 2011), in that they are made up of time-ordered visual 'slices' that convey timbral features of small windows of time from the original signal. This involves a process of overlapped windowing and down-sampling, as shown in figure 3.9.

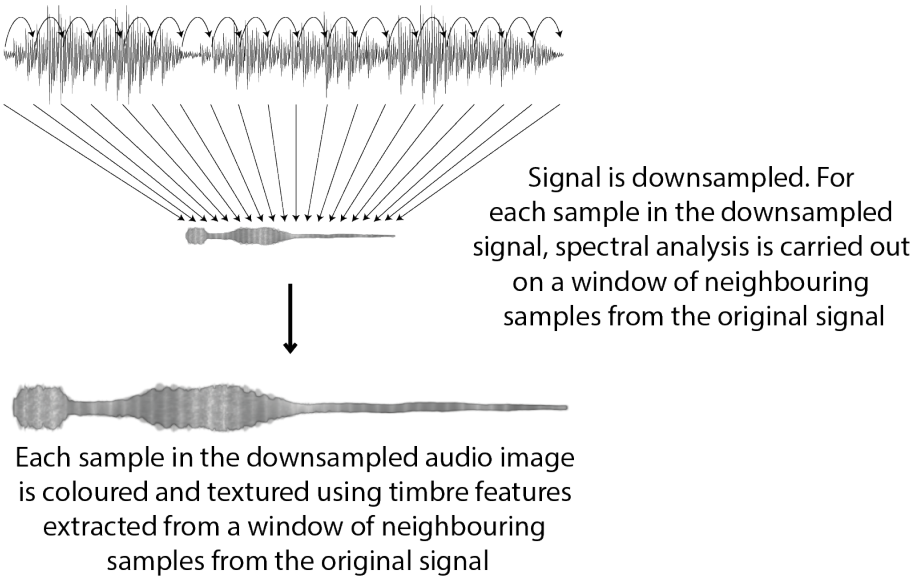


Figure 3.9: Downsampling.

The number of samples in the visual downsampled signal is set to 1024 such that each Sound Signature produces an image of 1024 x 128. The step size in the overlapped windowing is therefore dependent on the total number of samples in the recorded signature. For example, a recorded signature of 48000 samples results in a step size of 48000 / 1024 = 46.867. The size of the analysis windows is fixed at 512 samples. Windows are constructed from the original signal, centred around each sample in the downsampled signal. This way, timbre features are provided that correspond to each sample in the downsampled signal that is used for visual rendering. Figure 3.10 (a) shows in detail how the windowing process is implemented.

Timbre Feature Extraction

Timbre features are extracted for each window, as shown in figure 3.10 (b). FFT analysis is then performed on the scaled audio window. From the real frequency data, the spectral characteristics are calculated. The three main timbre features calculated are spectral centroid, spectral flatness, and spectral spread. These features give a good description of the spectrum of a short-time audio window (Peeters et al. 2011). The spectral centroid indicates the spectral 'centre of gravity' (i.e. where most of the energy is centred in the spectrum). The spectral flatness roughly indicates how noisy versus how tonal the audio is. The spectral spread measures how 'spread out' the spectrum is, around the centroid value. The features are summarised in table 3.6, along with suggested semantic mappings from existing studies. Table 1.1 provides definitions for the various terms used in the mathematical definitions of the features.

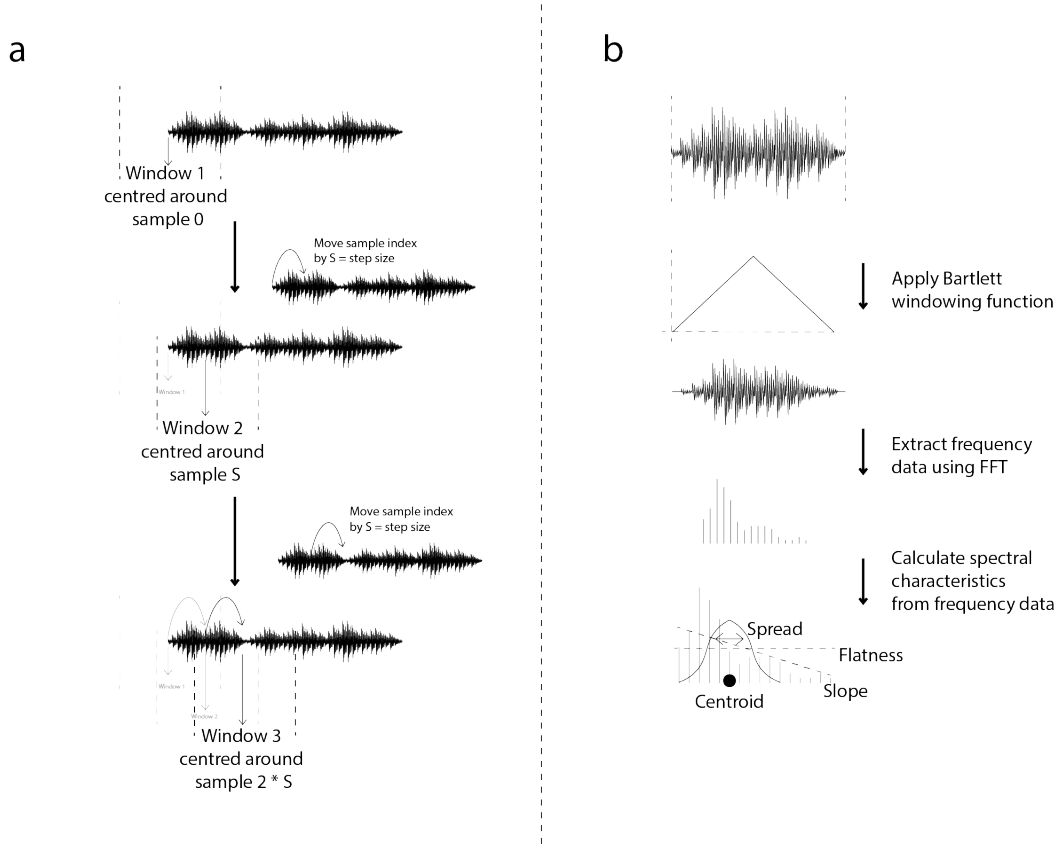


Figure 3.10: (a) Windowing and overlapping. (b) Feature Extraction

Table 3.6: Acoustic Timbre Features and Semantic Mappings. (See table 1.1 for mathematical key).

Acoustic Feature	**Mathematical Definition**	**Suggested Semantic Mapping(s)**
Spectral Centroid	$\mu_1(t_m) = \sum_{k=1}^{K} f_k \cdot p_k(t_m)$	**Brightness** Beauchamp 1982 De Poli and Prandoni 1997 Berthaut et al. 2010 Zacharakis et al. 2014
Spectral Spread	$\mu_2(t_m) = \left(\sum_{k=1}^{K}(f_k - \mu_1(t_m))^2 \cdot p_k(t_m)\right)^{1/2}$	**Texture Contrast** Giannakis and Smith 2000
Spectral Flatness	$SFM(tm) = \frac{\left(\prod_{k=1}^{K} a_k(t_m)\right)^{1/K}}{\frac{1}{K}\sum_{k=1}^{K} a_k(t_m)}$	**Texture Granularity** Giannakis 2001 Berthaut et al. 2010

Rendering

A large concatenated feature buffer is constructed that contains the timbre features for all of the samples in the downsampled signal. This feature buffer contains the centroid, flatness, and spread for sample 1 followed by the centroid, flatness, and spread for sample 2, and so on until sample N = the number of samples in the original signature. This buffer is wrapped into a 2D texture and sent to the GPU such that features can be easily accessed within a fragment shader. The 2D texture is structured such that the individual feature values for all samples are arranged accross each row.

The colouring algorithm is implemented in a fragment shader. This algorithm describes the colouring of a pixel based on its position and the corresponding audio features for that position. Details of the various timbre-visual mappings are discussed in the next section.

3.3.2 Visual Mappings

Colours are constructed in HSV space and the quantities for hue, saturation and value are determined by various timbre features. As mentioned previously in section 2.4.2, existing research has highlighted some correlations between acoustic features of timbre and semantic descriptors of timbre. One such correlation which has been identified in numerous studies is the correlation between the spectral centroid and the 'brightness' of a sound (Beauchamp 1982; De Poli and Prandoni 1997; Berthaut, Desainte-Catherine, and Hachet 2010; Zacharakis, Pastiadis, and Reiss 2014).

The 'value' quantity in the HSV colour space is therefore dependent on the spectral centroid value (figure 3.11).

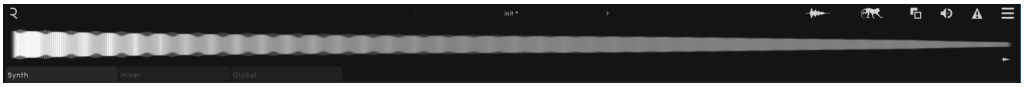


Figure 3.11: Spectral centroid mapped to brightness for a plucked string sound.

Correlation has also been identified between the noisiness of a signal and visual texture granularity (Giannakis 2001; Berthaut, Desainte-Catherine, and Hachet 2010). This mapping is intuitive to understand since auditory noise and visual noise mean very similar things. A noise texturing function is therefore implemented in the colouring algorithm, and the degree to which this affects a pixel's colour is dependent on the spectral flatness value (figure 3.12).

Figure 3.12: Spectral flatness mapped to noise amount. The sound in this example starts with a burst of noise, which is then filtered out over time.

Sound Signatures also make use of colour hue in order to depict frequency content in a similar way to the Comparisonics waveform display. The spectral centroid measures the 'centre of gravity' of the spectrum and can therefore be used as an indication of whether there is more low, mid, or high frequency in the signal. The spectral centroid value can be mapped to the hue value either directly or inversely in order to produce a red-to-blue or blue-to-red visual mapping (respectively) for low-to-high frequencies. Arguments can be made for both mapping strategies, and a user study has been set up in order to test whether there is any agreed user-preference for one over the other. This is detailed in the next section. Based on the use of the hue value to visualise the spectral energy content, the spectral spread value is represented using the colour saturation value. This way, the centroid value determines whereabouts to sample the hue-wheel, and the spectral spread value indicates how concentrated that colour should be. The spectral spread is mapped to colour saturation inversely, such that a low spectral spread leads to a more saturated colour (figure 3.13). Colours with higher saturation values will produce more contrast in the resulting texture. An association between spectral spread ('compactness') and texture contrast was highlighted by Giannakis and Smith (2000).

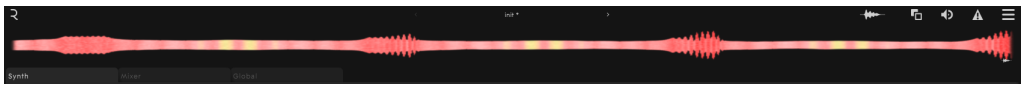


Figure 3.13: Spectral spread inversely mapped to colour saturation. The sound in this example has a bandpass filter applied, and the bandwidth of the filter changes smoothly between wide and narrow over time.

Once the colour has been constructed in HSV space using the fragment shader, a secondary fragment shader is used in order to produce a blurring or 'smoothing' effect. Prevalence of low-frequency energy over high-frequency energy leads to more pronounced blurring, such that high-

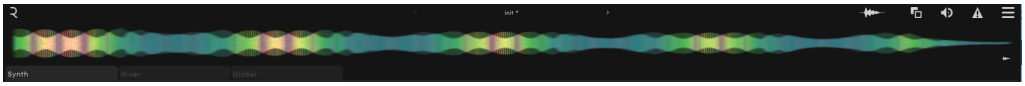


Figure 3.14: Spectral centroid controls blurring amount. The sound in this example has a low-pass filter, The cutoff rises and falls over time.

frequency content leads to more distinguished detail in the waveform (figure 3.14). In order to achieve this in the smoothing shader, the colour from the initial fragment shader is converted back to HSV space, and the 'value' quantity is used to determine how strong the blurring effect should be. This mapping is implemented inversely, such that a high brightness value produces less blurring. The use of the value quantity to determine the level of blurring assumes that the spectral centroid has been mapped to 'value' in the initial colouring stage.

3.3.3 Sound Signature Survey

An interactive survey was designed in order to examine participant preferences for different visual mappings of timbre features in the context of Sound Signatures. As discussed in the rest of this section, the interactive survey demonstrates some methodology that has been limited in other such studies. This includes the consideration of temporal variation of spectral features, animated visual stimuli, and a 'control choice' in the visual stimuli.

The survey was implemented as an interactive web app such that a maximum number of participants could be reached. The web app can run in a standard web browser and involves audio stimuli for which there are multiple visual stimuli (Sound Signatures) to choose from.

3.3.4 Stimuli

Audio Stimuli

Audio stimuli were chosen from a set of 100. All of the audio stimuli consist of audio clips that were generated using the ROLI Equator synth. Some featured only wavetable synthesis, some featured only samples, and some featured both wavetable synthesis and samples. Each audio clip was less than one second in length. Each consisted of one sustained note. Some had sustained temporal modulation such as an LFO controlling a low-pass filter cutoff frequency. They were generated using the various presets included in the Equator synth. One of the main goals was to capture a variation of timbres that may be found in such digital tools. For each audio clip, the audio data was used to generate four different Sound Signatures that served as the visual stimuli in the survey.

Visual Stimuli

Four 'visualisation strategies' were used in the survey, as shown in figure 3.15. Each mapping strategy uses various visual mappings for different timbre features. The first, referred to as *amp*,

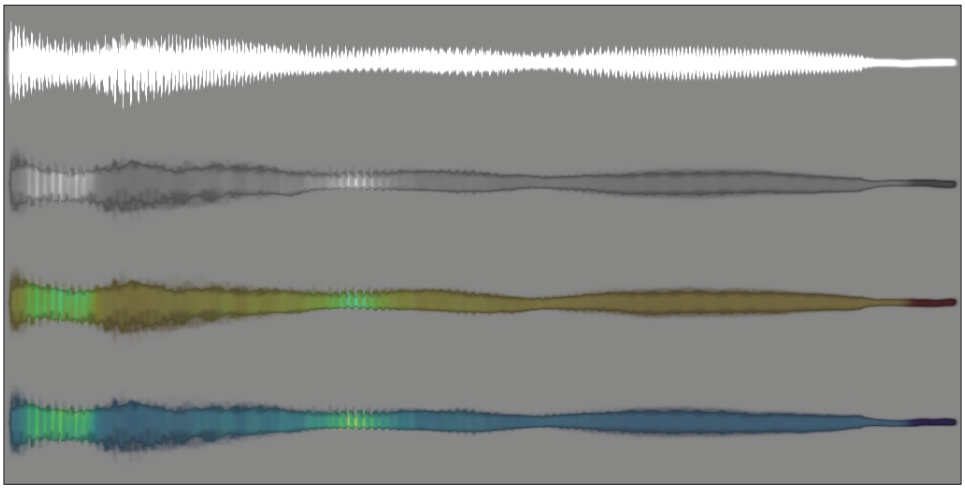


Figure 3.15: Examples of the different visualisation strategies. Top to bottom: *amp, grayscale, col, col-inv.*

is a basic amplitude waveform, with the exception that it visualises panning information using the vertical axis. Content above the centre line represents the absolute amplitude values of the left channel in a stereo signal. Similarly, content below the centre line represents the right channel. The second visualisation strategy - *grayscale* - maps spectral centroid to HSV value (brightness). It also maps spectral flatness to noise texture amount and maps the spectral centroid value to the amount of blurring in the effects shader pass. The third and fourth visualisation strategies (*col* and *col-inv*) use all of the previously discussed mappings with the additional mapping of spectral centroid to HSV hue and spectral spread to colour saturation (inversely). The third strategy - *col* - maps spectral centroid to HSV hue directly whereas the fourth strategy - *col-inv* - maps spectral centroid to HSV hue inversely. Thus, the *col* strategy produces reds-to-blues for low-to-high centroid values and the *col-inv* strategy produces blues-to-reds for low-to-high centroid values.

The *amp* visualisation strategy is intended to act as a control in the experiments. Forced-choice experiments often make the assumption that one choice is appropriate when in fact none may be appropriate. The *amp* strategy gives the option of indicating that none of the coloured or textured visualisations were preferred over the standard amplitude waveform.

3.3.5 Participants

93 participants took part in the study (11 female). At the beginning of the survey, participants are asked to fill out a short autobiographical questionnaire. This includes age, gender, musical experience (composition and performance), and musical production experience. For musical experience and musical production experience, participants choose from 5 options: none, amateur, intermediate, advanced, or professional. Figure 3.16 shows the proportion of participants that reported each level of experience (for both musical and musical production experience).

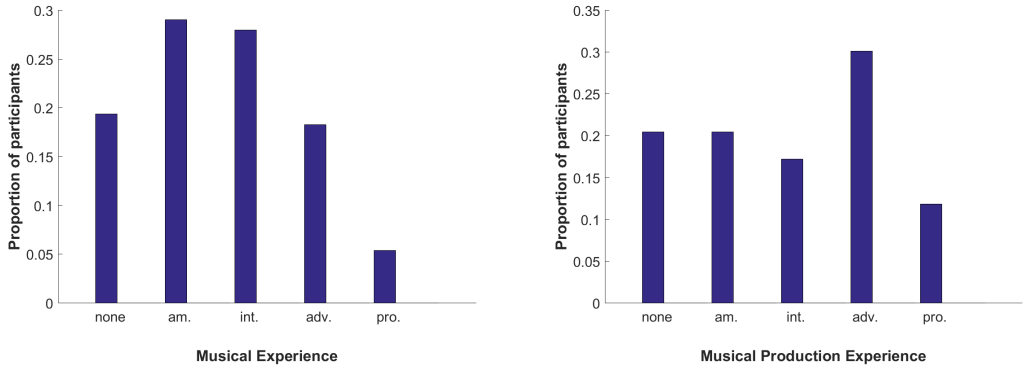


Figure 3.16: **Left:** Proportion of participants for each reported level of musical experience. **Right:** Proportion of participants for each reported level of musical production experience.

3.3.6 Experimental Procedure

In the survey, participants are asked to indicate which visualisation strategy they prefer for a number of different sounds. In each example there is one audio sample and four different visualisation strategies to choose from (*amp, grayscale, col,* or *col-inv* as described above). The visualisations are presented in a random order for each example. The audio can be previewed using the 'play audio' button. Whenever the audio is previewed, a playhead is animated from left to right across all of the visualisations. This animation is intended to emphasise the mappings between instantaneous spectral features (e.g. centroid) and visual features, and to draw the participants' attention to the temporal variation of audio features and the left-to-right timeline structure of the visual stimuli. When the participant has decided which visualisation they prefer, they click directly on that visualisation. They are then taken to the next example.

3.3.7 Data Collection

Data is recorded on a 'per-answer' basis as participants progress through the survey. Information is recorded every time a participant clicks on a sound signature option. This is implemented in order to ensure that missing data (e.g. from an unfinished survey) doesn't have an adverse effect on the results analysis. For each question that is answered, a data item is sent to an online database including the participant's biographical information, the question number, and their preferred visualisation strategy for that question.

3.3.8 Results

Figure 3.18 (**left**) shows the proportion of examples for which each visualisation strategy was preferred. Figure 3.18 (**right**) shows the proportion of users who preferred each visualisation strategy. In this context the preferred visualisation strategy is that which the user chose most

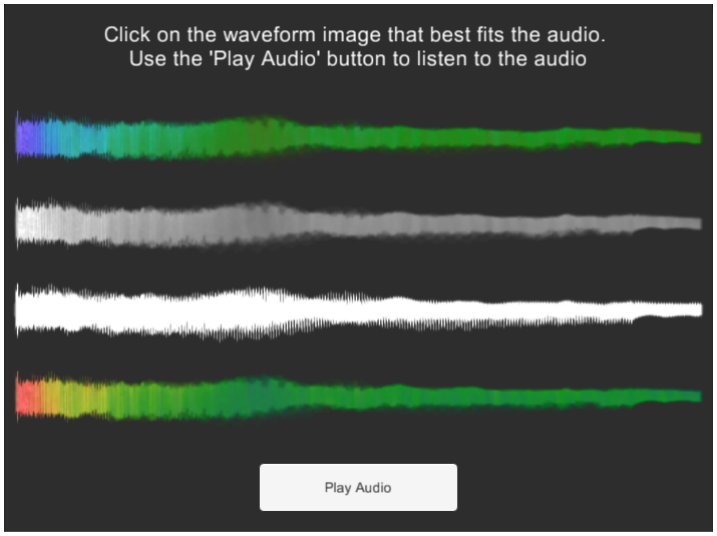


Figure 3.17: Survey Question Example

often. Chi square tests can be used to test for relationships between two categorical variables. Chi square tests for independence were performed in order to examine dependency between preferred visualisation strategy and other variables. Table 3.7 shows the p values for different tests of independence between preferred visualisation strategy and other variables. Figure 3.19 shows the proportion of examples for each level of musical (**left**) and musical production (**right**) experience for which each visualisation strategy was preferred.

Table 3.7: Dependency between preferred visualisation strategy and user information

Independent Variable	**p Value**
$Gender$	0.0701
$MusicalExperience$	0.0114
$MusicalProductionExperience$	0.0142

3.3.9 Discussion

The presence of 'control' visualisations (the *amp* visualisations) in the user study, combined with the fact that preferences were higher for the other visualisation types shows that participants preferred the use of Sound Signature representations over standard amplitude waveform representations.

The results of this survey show a clear preference for the *col-inv* visualisation strategy among participants. Adeli, Rouat, and Molotchnikoff (2014) showed that participants showed a tendency to pair 'soft' timbres with blues and greens and 'harsh' timbres with reds and yellows. The preference for the *col-inv* strategy is in agreement with this result, and quantifies the scale from 'soft' to 'harsh' as a scale from low spectral centroid to high spectral centroid value. There are

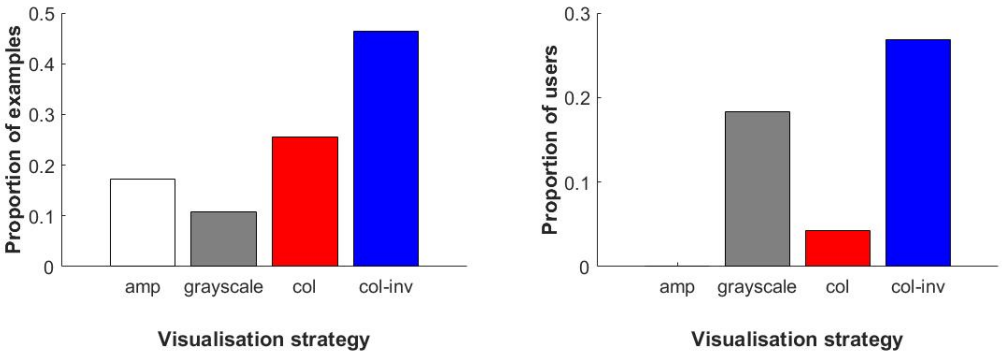


Figure 3.18: **Left:** Proportion of examples for which each visualisation strategy was preferred. **Right:** Proportion of users for which each visualisation strategy was preferred.

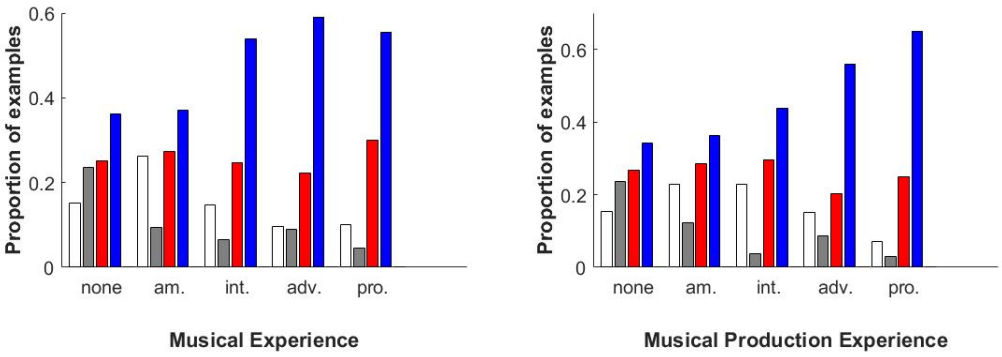


Figure 3.19: **Left:** Proportion of examples per level of musical experience for which each visualisation strategy was preferred. **Right:** Proportion of examples per level of musical production experience for which each visualisation strategy was preferred.

very few - if any - other studies that have quantitatively evaluated user preferences for spectral centroid-colour mappings in this way.

The results also suggest that preferred visualisation strategy is dependent on both level of musical experience and level of musical production experience, both at the 95% confidence interval (p value < 0.05). Figure 3.19 shows that the preference for the *col* and *col-inv* strategies was stronger for the more experienced participant groups (both musical and musical production experience). This disagrees with the results found by Adeli, Rouat, and Molotchnikoff (2014) who found no significantly unique clusters of user demographics after running a clustering analysis algorithm on user responses. This might be due to the fact that the visual stimuli used by Adeli et al. were very abstract in nature, consisting of monochrome complex 2D shapes. It is therefore unlikely that any of the participants in their study had previous experience in pairing those kinds of visual stimuli with audio stimuli. Sound Signature images on the other hand are very similar to amplitude waveforms, which musicians and music production practitioners are usually familiar with. This results in a higher probability of that portion of the participant pool having previous experience linking audio and visual stimuli similar to those used in the experiment.

Sound Signature representations have potential applications in numerous sound editing and sound design contexts. As discussed earlier in this section, the use of standard amplitude waveforms for audio representation has been shown to be insufficient for numerous audio production tasks. In digital audio workstations, amplitude waveforms are used to represent multiple stacked audio tracks. As is the case for Comparisonics coloured waveform displays, Sound Signature images could be used in this context to highlight specific timbral qualities of different tracks. Similarly, by highlighting regions of differing timbral qualities within individual tracks, Sound Signature images could be used to simplify the process of audio editing and region selection. Sound Signature images could also be used to streamline the search and retrieval process when searching through large libraries of audio samples. One potential area of application of the Sound Signature technique is visual scanning, or browsing, of audio samples in a library. It would be interesting to test whether the presence of Sound Signature representations has any effect on the efficiency with which users are able to select a desired or 'target' sample from a list.

3.3.10 Conclusion

This section has described the implementation and evaluation of a novel audio representation technique known as the Sound Signature technique. The Sound Signature technique follows the principles of a spectrogram representation in that the spectral content of small windows of time is presented using slices of a time-ordered visual depiction. Like the Comparisonics waveform display technique, Sound Signature images combine this spectrogram-influenced visualisation method with a standard amplitude waveform in order to combine the representation of dynamics and timbre.

An online user survey has been described which was implemented in order to investigate participant preferences for Sound Signature visualisations and visual mappings. The main objective was to identify if there was preference for the Sound Signature visualisations over standard amplitude waveform visualisations. Results suggest that there was preference for the Sound Signature visualisations. A secondary objective was to identify participant preferences for specific visual mappings in the context of Sound Signature visualisations. Again, the results show that there is a significant preference for specific mappings over others. Participants preferred the use of colour hue combined with brightness to represent spectral centroid over only the use of brightness. Participants also

preferred the use of an inverse mapping between spectral centroid and colour hue over the use of a direct mapping.

The study was implemented as a browser-based application such that a large participant population could be reached. The use of a browser-based survey also means the survey was taken in multiple different environments under various different conditions. The large participant population and the presence of clear preferences despite the various conditions give the results a higher degree of generalisability than other lab studies with smaller participant pools. Future studies involving the Sound Signature technique should investigate its use in specific contexts and compare the efficiency of participants when completing goal-oriented tasks on audio samples or synthesis presets both with and without Sound Signature representations.

3.4 Conclusion & Summary

This chapter has presented experiments and experimental tools that were developed in order to investigate mappings from timbre features to visual qualities. The identification of such mappings is important in the development of a perceptually motivated approach to timbre representation.

A preliminary study investigated the use of 3D animated visual stimuli in order to represent timbre. The user study identified clear preferred visual mappings when timbre features were varied in isolation. When multiple features were varied in parallel, there was no clear preference. Individual participant preferences for visual mappings of timbre features also varied in the presence of concurrently varying timbre features. The participant pool was very small and the level of sound technology experience was fairly limited within the participants. However, graphical representation techniques were explored that facilitate the visualisation of timbre as a virtual 3D object with descriptive structural and textural qualities. These techniques form the basis for a system described later in the thesis.

In order to address the issue of interface design and representation of a complex parameter space in the ROLI Equator software, a tool was developed in order to facilitate complex mappings from synthesis parameters to parameters of a visual effects engine. This allows the configuration of the Equator engine to be represented perceptually with reference to metaphorical visual qualities, rather than technical labels. This technique has been termed ‘synthesis visualisation’. The visual synthesis representations that were explored have some parallels with the visual sound shapes of spectromorphology presented by M. Blackburn (2011). They are abstract and illustrative. In a similar way to the visual sound sculpting proposed by M. Blackburn (2011), visual synthesis representations can be used for selection of presets. They can then be combined and altered in a process of visual synthesis sculpting.

The technique of synthesis visualisation has similar drawbacks to timbre notation techniques from electroacoustic music. It is abstract, subjective, and context-dependent. A different technique was therefore developed that makes direct use of low-level acoustic timbre descriptors. This is referred to as the ‘Sound Signature’ technique. It is similar to other data-driven temporal timbre representation methods described in section 2.5.3. The mappings used in Sound Signature representations are influenced by results from previous studies investigating visual timbre representation. An online interactive survey was conducted in order to evaluate the mappings used. The results of the survey showed that Sound Signature representations are generally preferred over

standard amplitude waveforms. The survey also showed preferences for an inverse mapping from spectral centroid to colour hue. The Sound Signature technique has potential applications in many types of application for sound editing and sound design. It forms the basis for an extension to the Equator interface that is described later in the thesis in section 4.2.

Chapter 4

Systems & Applications

The experimentation and experimental development described in the previous chapter was focused on the identification of mapping strategies between timbre features and visual features in different contexts. This chapter describes various systems that have been developed in which such mapping strategies are implemented and used as the basis for representation and interaction. This chapter also describes numerous audio-visual performances and installations in which timbre-visual mapping strategies were used to generate procedural visuals from musical performances.

Building on some of the key findings from the experimentation described in the previous chapter, the systems and performances described in this chapter were developed in order to provide demonstrative solutions to the research issues and questions being addressed in the thesis. All of the timbre representation techniques implemented in these systems and performances make use of perceptually motivated representation of timbre. Two of the systems were developed as extensions to Equator, an existing digital interface that uses a standard, engineering-focused layout with the usual technically-labeled controls found in most interfaces of this kind. The systems demonstrate how the perception and semantic description of timbre can be used to influence the design of digital interfaces for timbre representation and manipulation. Similarly, the performances described in this chapter demonstrate how the perception and semantic description of timbre can be used to influence timbre representation in a real-time performance context.

The first section describes a standalone real-time feature extraction application that was built in order to facilitate both the development of novel systems and the implementation of real-time timbre-driven visuals in live music performance contexts. The second section describes the 'Sound Signature' concept and an extension to the Equator interface involving coloured amplitude waveform visualisation of temporal timbre variation. The novel representation technique is described in terms of its integration with the Equator interface and the new interactive qualities it affords. The Sound Signature technique demonstrates how semantic timbre representation can be put to use in digital sound tools. The third section describes a real-time timbre manipulation system that represents Equator's audio output as a virtual animated 3D object. Gestural interaction within the object then controls the synthesis parameters in Equator. Again, this system demonstrates how semantic timbre visualisation can influence the design of digital interfaces. It also shows how semantic timbre visualisation can be used in a performative setting. The fourth section describes

various audio-visual performances that were produced in collaboration with a London-based arts collective. The events were all focused on the real-time visualisation of musical timbre during a live concert. They show how semantic timbre visualisation can be used in such contexts to draw attention to timbre variation as a performative and compositional device. The final section describes the development of a system that was influenced by the various audio-visual performances. The system represents timbre in real time using a fluid simulation algorithm.

4.1 Feature Extractor Application

A dedicated timbre-feature extraction application (henceforth referred to as Feature-Extractor) was developed for use in the demonstration systems described in this chapter. It calculates all of the features described in section 1.5 in real-time. The features are re-calculated continuously for a given audio stream, providing real-time tracking of the spectral and harmonic evolution of the audio. The features can then be sent via OSC such that other applications can use them for visualisation. The Feature-Extractor application can be used to analyse either an audio file as it plays in real-time or a real-time audio input signal (e.g. from a microphone). The application can analyse the left or right channels individually, or the entire stereo signal.

Audio is windowed using Bartlett windowing, with 256 samples per window for spectral analysis and 2048 samples per window for harmonic analysis. Therefore, at a sample rate of 48000 Hz, the total frequency bandwidth covered is 24000 Hz and this is split into 128 real-valued frequency bins for spectral analysis and 1024 real-valued frequency bins for harmonic analysis. From the 128-bin spectral envelope, four key features are extracted: spectral centroid, spectral slope, spectral spread, and spectral flatness. From the 1024-bin envelope, three key features are extracted: F0 estimation, harmonic energy ratio, and inharmonicity. See section 1.5 for definitions and descriptions of these features.

Feature-Extractor was developed in C++ using the JUCE audio development library. The JUCE library is an excellent library for developing audio applications and is one of the industry standards for the development of audio VST plug-ins. JUCE enables the development of fully cross-platform applications and Feature-Extractor runs on both Windows and Mac OSX.

The application is intended for real-time contexts and as such is designed to run as efficiently as possible. The OSC functionality is quickly and directly configurable such that the audio features can be used as control parameters for an external system within seconds of opening the application.

Stark (2014) developed the 'Sound Analyser' plugin-in - a very similar system to Feature-Extractor, which is intended for use in similar contexts. The 'Sound Analyser' provides access to a rich feature set which is more extensive than that of Feature-Extractor. It also provides direct access to OSC functionality such that the features can quickly be used in other applications. Sound Analyser is written as a plug-in, which means it runs in some kind of host such as Logic Pro or Ableton. This brings many benefits including multi-track analysis, simplified pre-processing (through the use of features provided by the host and other plugins), and automatic saving and recalling of analysis steps and parameters by the host. The main reasons why Sound Analyser was not used for the projects detailed in this chapter involve timing and constraints. The initial intention behind the Feature-Extractor application was to isolate the feature extraction code that had been integrated into the ROLI Equator sound engine and build a much more light-weight

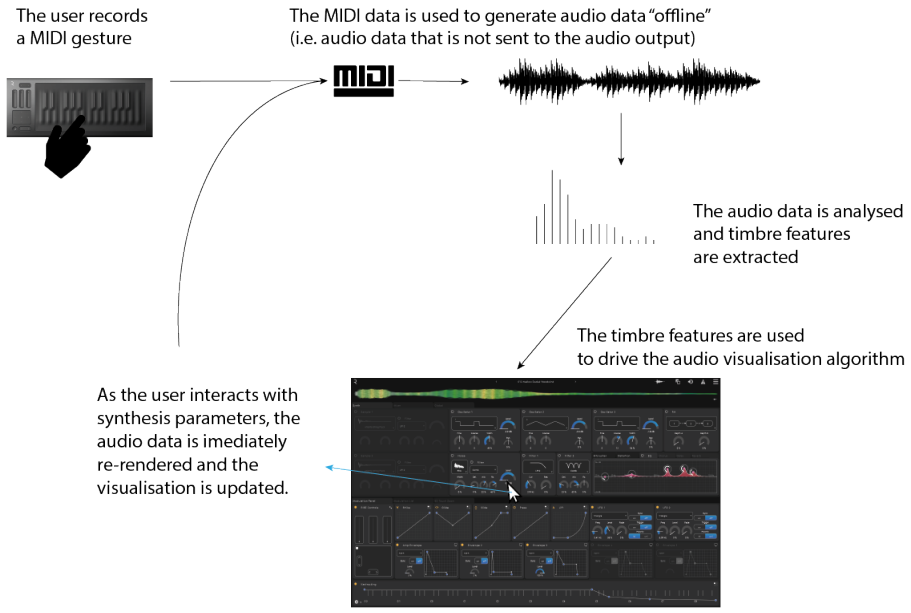


Figure 4.1: Sound Signature Overview.

application that could be used for continued research in other contexts. At the time Sound Analyser was presented, Feature-Extractor had already been developed and was in use. It evolved in response to project needs and therefore became very suited to the continued research.

4.2 Sound Signature

In the previous chapter, an experiment was described that made use of a novel timbre representation technique known as the Sound Signature technique. This section describes the development of an extension to the Equator interface based around the Sound Signature technique. The Sound Signature visualisation tool consists of a coloured amplitude waveform display. This waveform represents audio that is generated by the Equator engine, from a MIDI touch gesture. The MIDI gesture can be recorded by the user or auto-generated. Figure 4.1 gives an overview of how the Sound Signature feature works within the Equator interface.

Firstly the user records a short single-note MIDI gesture. For example they might press a note, apply some aftertouch (pressure) to the note, and then release the note. This MIDI data is then stored as a 'MIDI gesture' such that it can be used to produce audio data whenever the state of the sound engine changes (e.g. when the user changes a parameter or switches presets). The Sound Signature tool has some unique features that have been implemented in order to maximise its usability in a sound-design context.

4.2.1 'Offline Rendering' & Direct Interaction

As the user tweaks synthesis parameters in the Equator interface, the Sound Signature image updates immediately. This functionality has been implemented in order to give the impression that the user is 'sculpting' the sound in real-time. In order to achieve this, a duplicate version of the Equator engine is used (henceforth referred to as the 'offline engine'). This offline engine only generates audio data, but does not copy the data to the audio output buffers. This way, as the user tweaks parameters, the offline engine can be used to very rapidly re-generate audio data from the recorded MIDI gesture. This audio data is analysed and timbre features are extracted that are used to drive the visualisation algorithm, and the Sound Signature image is updated. This continuous updating of the generated audio data in real-time allows the Sound Signature to respond very quickly as the user manipulates the parameters of the engine. The effect of changing the filter cut-off, for example, is immediately visualised in the Sound Signature, giving an immediate indication of how the filter cutoff parameter affects the sound.

One thing that can be confusing for new users of sound design interfaces is that the effect of various parameters on the resulting sound *changes* depending on other configurations, and certain parameters have different meanings in different contexts. For example, the filter cut-off value has very different effects on the sound depending on whether a high-pass or low-pass filter is being used. When the sound signature is present, the act of changing the filter type will immediately produce visible results in the signature, which indicates how the two filter types affect the sound differently. This is then reinforced as the user tweaks the filter cut-off value and the signature updates accordingly in real time. This reduces the need for continued aural previews of the sound as the user tries to grasp what the parameter is doing.

This kind of immediate real-time feedback is achieved though the use of the novel offline rendering technique. In other words, the use of dual instances of the audio engine: one to produce audio at audio-rate in response to real-time performance input, and one to render audio data at faster-than-audio-rate from pre-recorded input data for the purposes of visual feedback. This technique is central to the Sound Signature concept and its integration with the Equator interface.

4.2.2 Sound Design Workflow

As mentioned earlier in this section, the generation of audio data from which the visualisation is rendered comes from the use of a recorded (or auto-generated) MIDI signature. This MIDI signature includes all of the input parameters such that they can be graphed over time and superimposed on the Sound Signature, as shown in figure 4.2. This indicates visually how the modulation parameter changes the audio over time. The user can also then assign that modulation parameter to different synthesis parameters and immediately inspect how it changes the sound. Equator's built-in modulators (i.e. LFOs and envelopes) can also be inspected in the same way (figure 4.3).

By visualising touch data and modulation data, and how this affects the sound over time, the Sound Signature can be used to present a visual snapshot of the parameter space and underlying timbre space that is afforded by the current configuration of the engine. For example imagine a signature that starts with harsh bright, noisy reds and then fades into a blurred blue, and then swells again to reveal clearer texture. This shows immediately a crisp, sharp attack to the sound followed by a dimmer, more dulled decay period, followed by a reintroduction of brighter content.

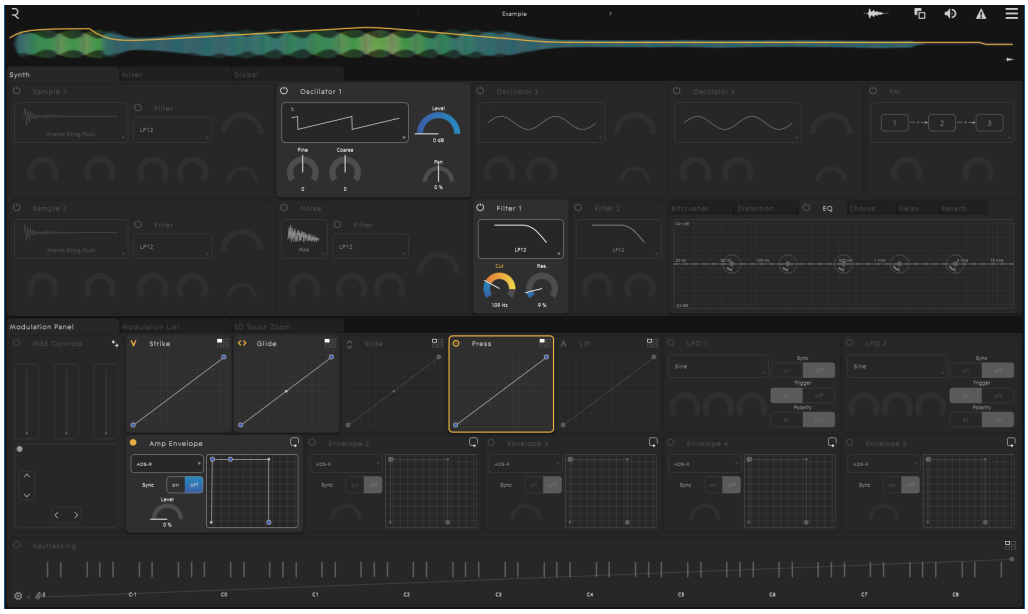


Figure 4.2: Modulation parameter inspection. Touch pressure modulates filter cutoff. The Sound Signature indicates how this affects the sound temporally as the pressure changes over time.

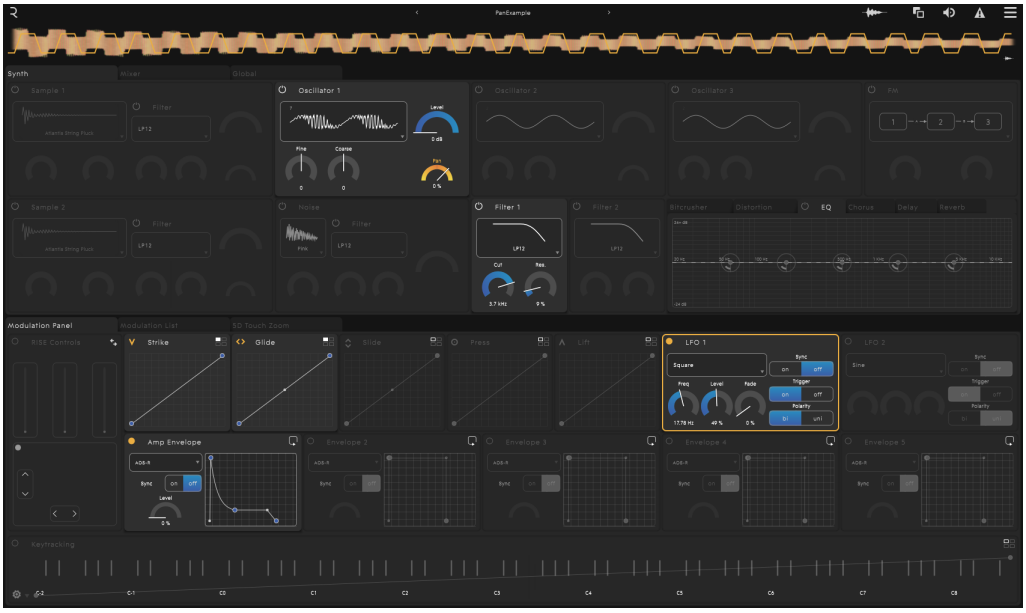


Figure 4.3: Modulation parameter inspection. LFO modulates pan. The Sound Signature indicates how this affects the sound temporally.

The user can then easily inspect the modulation parameters to identify what causes the evolution of the sound parameters.

Panning

Sound Signatures are not standard amplitude waveforms in that they are designed to visualise both the left and right channel of a stereo signal together, in one waveform. In order to achieve this, the absolute amplitude values are used. In a Sound Signature, colour above the centre line represents average absolute amplitude data from the left channel and colour below the centre line represents average absolute amplitude data from the right channel. This way, sounds with panning effects can be instantly identified, as shown in figure 4.3.

4.2.3 Semantic Timbre Description

Sound Signatures use texture, colour and brightness to visualise timbre features. As discussed in the previous chapter, the mappings are based on semantic descriptors of timbre and their acoustic correlates. In this way, common visual descriptors of sound qualities such as 'smooth / rough' or 'bright / dull' are used directly to visualise the relevant timbre features.

Consider the example of a low-pass filter being applied to a square wave. If the cut-off value is initially high, the sound signature will be bright (as the high centroid value produces high HSV value) and clear (as the high centroid value produces no blurring in the visualisation). If the user then reduces the cut-off value they will see the signature becoming duller and more blurred. This effectively bridges the gap between task language and core language as the user can see that reducing the cut-off effectively 'dulls' the sound.

The colour produced by a square wave will have lower saturation than that of, say, a sinusoidal wave, since a square wave produces wider spread in the spectrum. Again this visualisation aids in the understanding of the differences in sound qualities between different waveforms. It is immediately clear that sinusoidal waves are more 'concentrated' in a particular area of the spectrum.

The effect of adding distortion, or noise, to the audio signal will be immediately represented as the addition of visual noise to the sound signature. Effectively the user is making the sound 'more grainy'. This combines with the idea of sound design workflow integration discussed previously in section 4.2.2. For example if the level of distortion is controlled by touch pressure then the sound signature timeline visualises how applying more pressure will bring out the 'graininess' in the sound.

4.2.4 Preset Browsing

As well as individual timbre variation within a given sound, Sound Signature images highlight contrasts between different sounds. When arranged in a list, for example, they clearly indicate the differing timbre characteristics between sounds (Figure 4.4). Since they represent temporal timbre variation, Sound Signature images can effectively represent the differences between similar sounds,

or similar instrument types. For example, a bass sound with heavy distortion will be identifiable as more fuzzy and brighter than a non-distorted, low-pass filtered bass sound.

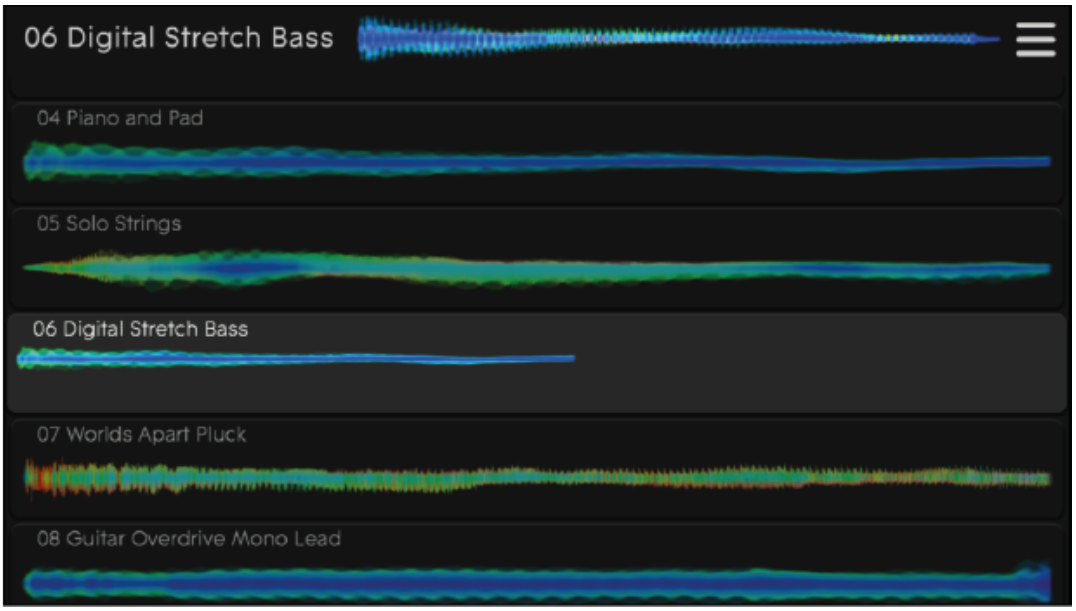


Figure 4.4: Sound Signature representations used in a preset browsing context.

4.2.5 Discussion

Sound Signature images represent temporal timbre variation by mapping spectral timbre features to visual properties of a waveform-like visual representation of the audio. As is the case for the techniques presented by Rice (2005) and Malt and Jourdan (2011), one of the major advantages of Sound Signature images is that they extend the traditional familiar amplitude waveform representation and add a representation of timbre data. The mappings used in Sound Signature images are influenced by findings from existing studies. They have been evaluated using a large user study investigating user preferences for different visual mappings within the specific context of Sound Signature images. This study identified specific preferences for Sound Signature mappings over standard amplitude waveform representations.

The Sound Signature tool was developed as an addition and extension to a standard sound design interface with a user-defined architecture. As an extension to the interface, it demonstrates how perceptually motivated visual timbre representations can be used to provide visual reference points in a sound design interface. The 'offline audio rendering' method facilitates immediate visual representation in response to changing audio parameters. For example sound design processes such as changing waveform type, filtering, or applying effects, all produce perceptually relevant changes in the Sound Signature image, indicating semantically how each process affects the audio. By overlaying graphs of synthesis parameters over the Sound Signature images, their effect on the timbre can be visually understood from the changes in visual qualities on the image. In this way, the Sound Signature representation technique provides a bridge between task language and core

language, as discussed by Seago, Holland, and Mulholland (2004).

4.3 TimbreSphere

The Sound Signature tool described in the previous section makes use of temporal timbre representation and is designed for the context of *sound design*. This section describes the development of a *performative* tool that makes use of real-time timbre representation, referred to as the TimbreSphere tool. The previous chapter gave details of a preliminary study involving 3D animated visual stimuli that were implemented using deformation of a spherical mesh. This graphical visualisation technique was developed and extended throughout the development of the TimbreSphere system. The aforementioned study featured digitally synthesised audio stimuli. Similarly, the Timbresphere tool was developed in the context of digital synthesis.

The TimbreSphere interface affords performative exploration of a synthesis parameter space which is guided by semantically motivated visualisation of the timbre. The system constitutes a 'simplified counterpart' (as discussed in section 1.2.2) to an underlying synthesis engine with user specified architecture. The underlying software is the Equator sound engine from ROLI which has multiple sound sources, filters and effects. This simplified counterpart facilitates the manipulation and exploration of specific (fixed architecture) configurations of the engine (i.e. presets) through performative gestural control. Since TimbreSphere is a tool designed for the performative exploration of a synthesis parameter space, it has been designed with monophonic timbres in mind. It has been tested and demonstrated using monophonic synthesised and sampled sounds and was developed with the intended use of temporal timbral variation of synthesised/sampled sounds, either in a performance or sound design context. The system makes use of real-time visualisation of acoustic timbre features, combined with gestural control of synthesis parameters. As a performative extension to a standard sound design interface, the system provides an example of perceptual timbre representation being put to use in both a performance and an interface design context.

An overview of the system is given, followed by a detailed description of the visualisation techniques it uses and the various timbre-visual mappings involved. This is followed by a discussion of the interface's affordances. Finally, the system is reviewed with reference to existing literature and other systems.

4.3.1 System Overview

The system makes use of a Microsoft Kinect 2.0 sensor in order to track the user's hands. The front-end of this system consists of a virtual 3D space within which the user's hands are represented, along with a 3D visualisation of the audio being generated by the sound engine. Real-time audio analysis functionality is built into the audio engine, which continuously extracts timbre descriptors from the audio that is generated. These timbre descriptors are used to drive the visualisation in the user-facing front end. In turn, the X, Y and Z distances between the user's hands are mapped to different parameters within the sound engine. The user can thus drive the synthesis engine through gestural motion of their hands within the space. This leads to both audible changes in the sound and corresponding visual changes to the visualised object. Figure 4.5 shows a flow diagram of the system.

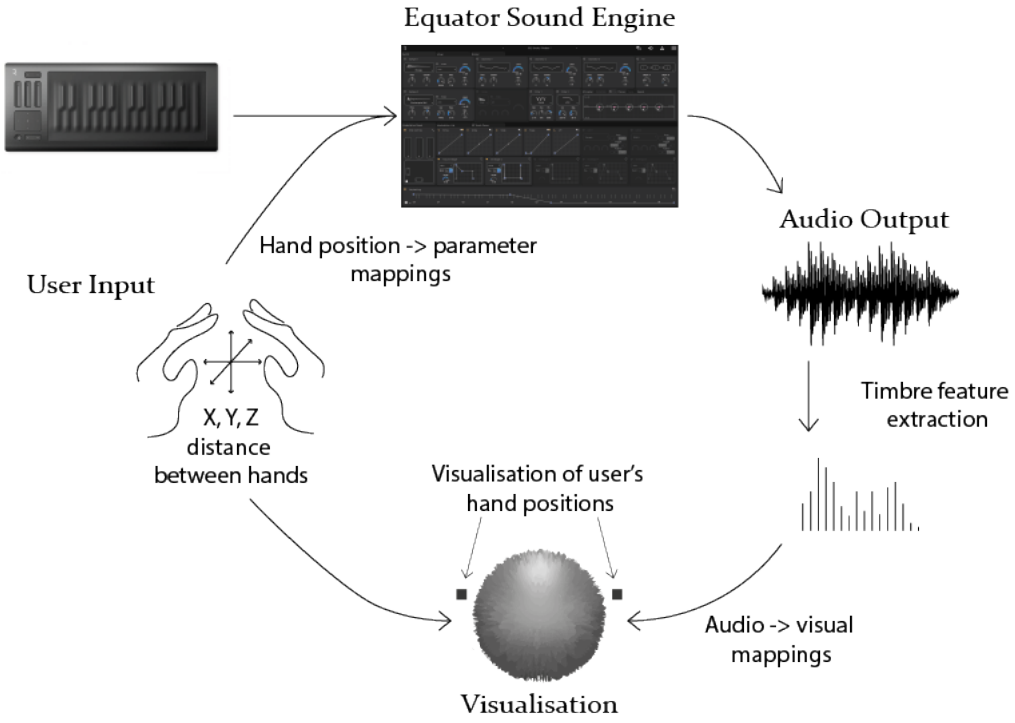


Figure 4.5: TimbreSphere system overview.

The rest of this section will detail each component of the system in detail.

4.3.2 Visualisation

The system visualises the audio being generated by the sound engine as a spherical structure with surface deformations, where the brightness, shape, and texture of the sphere are used to indicate different semantic descriptors of the timbre.

Luminance / Brightness

The sphere is rendered using Blinn-Phong shading (Blinn 1977). The brightness of both specular and diffuse components can be controlled individually. A third parameter is the specular radius (the radius of the specular highlight). Figure 4.6 shows the effect of varying the specular radius value.

A small specular radius leads to local specular highlights, whereas a large specular radius leads to widespread illumination.

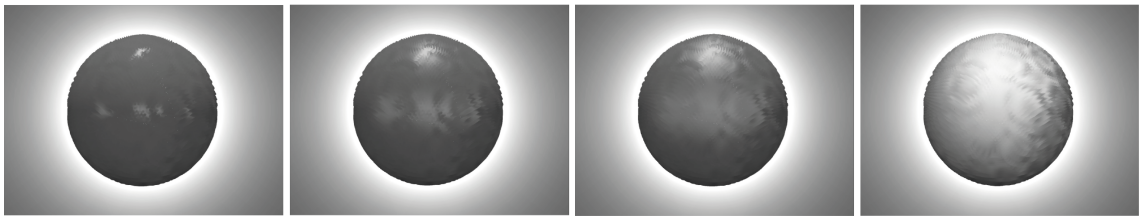


Figure 4.6: Varying specular radius value from low (left) to high (right).

Texture

The vertices of a spherical model each have normals pointing away from the centre of the sphere. Each vertex can therefore be perturbed along its normal vector in order to deform the surface of the sphere. The position of vertices on the surface of sphere can be defined by their azimuth angle (a_{az}) and their inclination angle (a_{inc}). In this system, a vertex shader is used in order to alter the position of vertices along their normals, depending on their azimuth and inclination positions on the sphere. Take the function

$$Dv_{az} = \sin(a_{az}). \tag{4.1}$$

This defines a deformation amount for each vertex, depending on the azimuth position of the vertex. By moving each vertex along its normal by the Dv_{az} amount defined by its azimuth position, a global distortion is applied to the sphere as shown in figure 4.7.

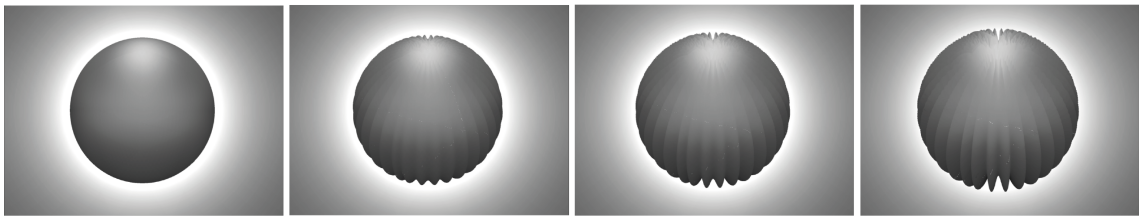


Figure 4.7: Spherical vertex extrusion dependent on azimuth position. Extrusion amount increases from left to right.

The frequency of deformation can be controlled by multiplying the azimuth position by a constant 'frequency' term (f_{az}):

$$Dv_{az} = \sin(a_{az} \cdot f_{az}). \tag{4.2}$$

A similar deformation amount can be set up using the inclination position:

$$Dv_{inc} = \sin(a_{inc} \cdot f_{inc}), \tag{4.3}$$

as shown in figure 4.8.

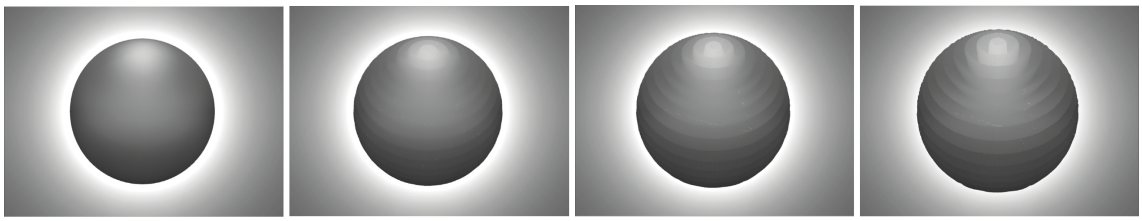


Figure 4.8: Spherical vertex extrusion dependent on inclination position. Extrusion amount increases from left to right.

The azimuth and inclination deformation amounts can be combined to define a unified deformation amount:

$$D_v = Dv_{az} + Dv_{inc}. \tag{4.4}$$

Figure 4.9 shows the effect of changing the azimuth and inclination deformation frequencies in parallel.

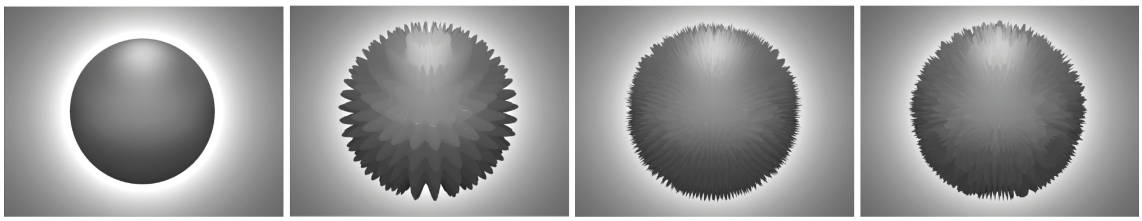


Figure 4.9: Spherical vertex extrusion dependent on azimuth and inclination position. Extrusion amount increases from left to right.

Three weighting terms E, e_{az} and e_{inc} can be added to control the magnitude of deformation:

$$D_v = E \cdot (Dv_{az} \cdot e_{az} + Dv_{inc} \cdot e_{inc}), \tag{4.5}$$

where E controls the overall deformation amount, e_{az} controls the deformation amount along the azimuth arc, and e_{inc} controls the deformation amount along the inclination arc. Altering the parameters f_{az}, f_{inc}, e_{az} and e_{inc} can lead to multiple different surface textures.

In addition to the vertex extrusion deformation, bump mapping is implemented by perturbing the normal vectors slightly, using a Perlin noise function. This noise function has two parameters: B_a the amount by which the normals are perturbed and B_g the granularity of the Perlin noise. This bump mapping technique produces the impression of coarse surface deformations on the sphere. Figures 4.10 and 4.11 show the effect of changing the amount and granularity, respectively.

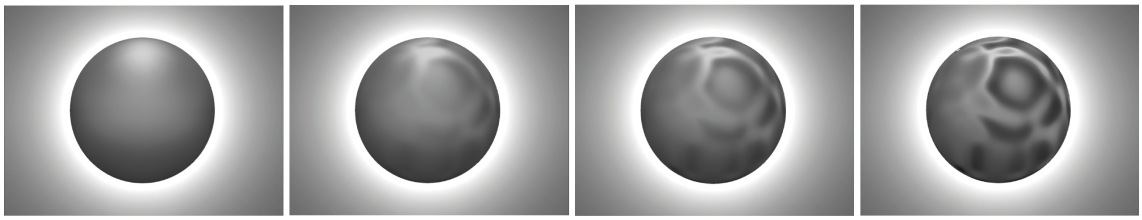


Figure 4.10: Perlin noise bump mapping with varying amount, from low (left) to high (right). Granularity remains fixed.

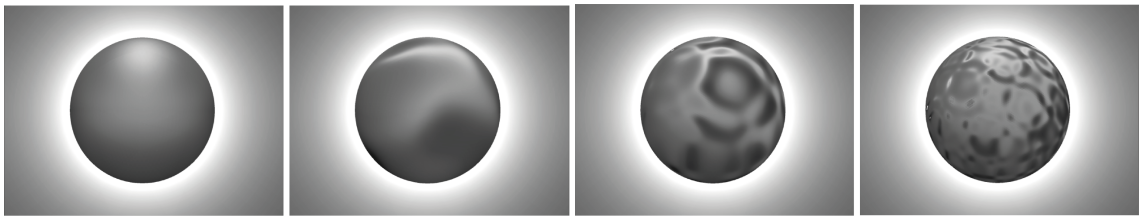


Figure 4.11: Perlin noise bump mapping with varying granularity, from low (left) to high (right). Amount remains fixed.

Mass / Volume

The volume of the sphere can be controlled parametrically. Additionally, the extrusion amount E which controls the amplitude of deformation around the sphere, can be altered.

4.3.3 Mappings

This section explains the mapping strategy that has been implemented in this system. It gives motivations for the mappings chosen with reference to existing research.

Luminance / Brightness

As mentioned previously, multiple studies have found that the spectral centroid provides a decent indication of the brightness of a sound (Beauchamp 1982; De Poli and Prandoni 1997; Schubert, Wolfe, and Tarnopolsky 2004; Schubert and Wolfe 2006). Therefore, the spectral centroid was used to drive the brightness in the visualisation. Additionally, the spectral spread is mapped to the specular radius in the visualisation. This way, more broadband audio will produce more evenly dispersed lighting, whereas narrow-band audio will produce particular highlights (figure 4.6).

Texture

The TimbreSphere system explores the use of spherical deformation to create an analogous 'waveform-like' shape around the surface of the sphere. The properties of this waveform deformation are driven by the spectral characteristics of an audio signal. The waveform-inspired texture is far from a direct representation of the audio signal, but instead is designed in order to display similar textural characteristics.

The spectral flatness measure is mapped to the frequency of inclination deformation in the vertex extrusion. The amount of inclination deformation is controlled by the spectral centroid. This combination of spectral flatness and spectral centroid mappings was designed in order to obtain a characterisation of the type of noise, as well as just an indication of the presence of noise. Noisier signals will produce rougher deformations on the sphere, and these deformations will be heightened when signals have a high centre of mass of spectral energy, distinguishing high-frequency piercing noise from low-level rumble. Links between texture roughness and noisiness have been shown by Giannakis (2001) and Berthaut, Desainte-Catherine, and Hachet (2010).

The frequency of the azimuth deformation is controlled by spectral centroid, and the azimuth amount is inversely controlled by the spectral spread. This means that when the spectrum is more grouped around the spectral centroid, there will be a noticable uniform deformation around the sphere, and the frequency of this deformation will increase with the spectral centroid.

The harmonic energy ratio is inversely mapped to the bump mapping amount, meaning that audio with less harmonic content will produce more pronounced deformations. Harmonic energy distribution was linked to texture by Zacharakis, Pastiadis, and Reiss (2014). The granularity of the bump mapping is controlled by inharmonicity, meaning that complex signals with inharmonic peaks (e.g. in the case of detuned sinusoid oscialltors) will produce more granular bump deformations (figure 4.11). Inharmonicity has been linked to texture-repetitiveness by Giannakis (2001).

Mass / Volume

The overall deformation amount in the vertex extrusion is controlled by the root-mean-square amplitude (RMS) of the audio, multiplied by the spectral centroid. Thus, louder audio will produce more pronounced deformations, with high centre of mass in the spectrum emphasising these deformations. A link between mass and spectral centroid was suggested in Zacharakis, Pastiadis, and Reiss (2014).

4.3.4 Gestural Control of Synthesis Parameters

The user's hands are tracked and visualised in virtual 3D space. The X, Y and Z distances between the hands are tracked and used to control different parameters within the Equator sound engine. The parameters that these distances control can be varied, and depend on the configuration of the sound engine for a given preset. The mapping of these distances to different synthesis parameters is what constitutes the creation of an exploration interface for a given fixed architecture configuration of the sound engine, in other words, the creation of a 'simplified counterpart' interface as discussed in section 1.2.2.

As an example, X distance might control filter 1 cutoff value, Y distance might control reverb amount, and Z distance might control delay time. The important feature of this system is the direct visual feedback provided by the perceptually motivated visualisation. As the user extends their hands along the x-axis, they hear the result of the filter opening, but also see a corresponding visual analogy. This provides visual context to the control, linking its aural affect with a visual analogy.

Rapid Prototyping of Gestural Interactions Using Equator

By mapping the X, Y and Z distance between the user's hands to different parameters in the sound engine, they effectively become control dimensions in a timbre space. Importantly however, they likely control multiple dimensions in the timbre space simultaneously, since parameters in the synthesis engine can change multiple timbre features of the audio simultaneously. Thus, exploration of the afforded timbre space is achieved through gestural, spatial exploration. Depending on the mapping between spatial dimensions and synthesis parameters, certain regions emerge in the physical space that produce certain unique timbral qualities. The idea is that the timbre space becomes realised in real physical space as a 3D area that can be explored. The method of exploration depends on the position of both hands, as their distance along each dimension is what drives the parameters.

As mentioned in section 1.5, the Equator interface is focused on modulation. It was designed with the intention of simplifying the process of linking modulation sources to synthesis parameters. The interface features three graphical sliders that are conventionally used to represent the different control sliders available on one of ROLI's hardware controllers, the Seaboard RISE. For the purposes of TimbreSphere, an OSC connection was developed such that the distances between the user's hands along the X, Y, and Z dimensions was directly mapped to these macro control sliders in the Equator interface. The Equator interface could then be used to easily map different dimensions in the TimbreSphere input space to various synthesis parameters. This allows the construction of complex n-m mappings between physical space dimensions and synthesis parameters.

Visual Sound Sculpting

The visualisation of the timbre as a 3D object that is being manipulated guides the exploration through the timbre space. Again, this kind of visualisation was intentionally designed as a bridge between 'core' and 'task' language. If the Y distance between hands is mapped to distortion level, for example, then the user sees the central sphere become rougher in texture as they increase the vertical distance between their hands. They learn that moving their hands to opposite vertical extremities produces a particular aural effect. Similarly, they can then combine this with aural effects produced by movement along other dimensions. They can build up a performative vocabulary and begin to explore the use of timbral variation in a performative context.

The 3D visualisation in the TimbreSphere system can be considered as a visual sound object in the same way that M. Blackburn (2011) talks of visual sound shapes in the context of spectromorphology. M. Blackburn (2011) proposes a form of visual sound sculpting through the use of perceptual visual illustrations of timbral qualities. This kind of visual sound sculpting is achieved in the TimbreSphere system by presenting the physical space in front of the user as a space in

which gestural interactions can be used to sculpt and manipulate the virtual object, simultaneously producing perceptually correlated timbral effects in the audio.

4.3.5 Discussion

The motivation behind the development of this system was the prevalence of textural, material and physical vocabulary involved in the semantic description of timbre. The intention was to make use of existing research into both acoustic descriptors of timbre and semantic descriptors of timbre in order to develop a novel interface for timbre manipulation. The interface was conceived as an extension to an existing interface, which has an engineering-focused parametric layout (the Equator synth). This extension provides a perceptually motivated front end to the sound engine, where, for example, 'rough' 'bright' sounds are presented as rough bright objects.

In the context of existing interface design, as discussed in section 1.2, this system has been designed in order to provide a performance-oriented interface. It was developed in order to demonstrate a novel form of exploration of a fixed-architecture parameter space, and the underlying timbre space. Navigation of the parameter space is semantically-driven, and involves parallel parameter control. It is guided by semantically motivated visualisation of the audio. This contrasts with the serial parameter-driven navigation afforded by the Equator interface. Linking gestural user input parameters to macro controls in the Equator interface, enables rapid prototyping of gestural control spaces for synthesis manipulation.

The visualisation system visualises the audio as a graphical 3D object with different surface texture, physical form, and luminance qualities. As mentioned in section 3.2, one of the main motivations for the choice of visualisation was its effectiveness in representing the notion of a sound source. This is particularly important in the context of digitally produced sound, where the physicality of the sound source is less important, and visual metaphors may be drawn upon to a larger extent. The 3D sound objects produced by the TimbreSphere system are similar to the visual sound shapes of spectromorphology discussed by M. Blackburn (2011) in that they provide perceptual representations of timbral properties. In the TimbreSphere system, gestural interaction in a physically realised timbre space enables a form of visual sound sculpting where these sound objects are manipulated and timbral effects are produced. This is similar to the process of visual sound sculpting described by M. Blackburn (2011) where spectromorphological sound shapes are used to illustrate timbral events.

Although this specific system makes use of a perceptually motivated timbre-visualisation framework for a performance use-case, such a framework would be of benefit in both sound design and search and retrieval use cases as well. In sound design contexts, it would aid in bridging the gap between task language and core language, as demonstrated by the Sound Signature tool. In search and retrieval contexts, it could be used to provide visual reference points to assist in the orientation of a search.

The TimbreSphere tool was developed in the context of performance and performative exploration of timbre spaces. However, it could be further developed and iterated to be more suited to a sound design context. The proof-of-concept TimbreSphere implementation that has been demonstrated makes use of the Kinect sensor for motion tracking. The use of something like a Leap Motion controller could be used for closer more detailed tracking of the hands. This kind of setup would be more harmonious with a standard mouse and keyboard setup as the user could

easily switch between the two. Gestural motion of the hands could then be used for a multitude of tasks such as audio scrubbing, modulation recording, envelope editing etc. Since the focus of the TimbreSphere demo, and of this thesis in general, is the use of semantic visualisations for timbre, these kinds of extensions were deemed out of scope for this project.

4.4 Analema Group Performances – Preliminary Case Studies

The previous sections have described the implementation of novel interface tools for timbre creation and manipulation that make use of perceptual timbre representation. The next two sections describe three audio-visual performances that were produced in collaboration with a London-based audio-visual arts collective, which explored the use of perceptual timbre representation in a real-time performance context.

The performances were all based around real-time tracking of audio features of live performers during a musical performance, using the Feature-Extractor application described in section 4.1. For each performance, visual mapping strategies were designed whereby timbre features from Feature-Extractor were mapped to visual parameters of a specific visualisation system. The development of these visual mapping strategies is described in this and the next section. Each performance is presented as a case study into the development of a perceptual timbre representation system for a specific context, with specific constraints. The first two performances (described in the rest of this section) were pilot case studies that informed the development and implementation of the third performance. The third performance (described in the next section) took place on a much larger scale and was more technically challenging. Evaluation data from both audience members and performers is presented and analysed for the third performance. The first two case studies are evaluated more in terms of how they influenced the implementation of the third performance.

This section is structured as follows. Firstly, a brief description is given of the Analema Group – the audio visual arts collective with whom the performances were collaboratively produced. Secondly, motivations are given for the use of stochastic motion simulators as visualisation systems in all of the performances. This is followed by descriptions of the two initial pilot-study performances that were implemented. Each performance is also discussed with relation to spectromorphological concepts, and spectromorphological representation.

4.4.1 Analema Group

Analema Group is an audio-visual arts collective based mainly in London. The key objective of Analema Group is to create unique transformative experiences that reflect on the nature of perceptual phenomena, exploring the relationships between sound, colour, light, movement and form.

4.4.2 Stochastic Motion Simulation for Real-Time Timbre Representation

During the performances various acoustic timbre features were mapped to visual parameters of visualisation systems. These visualisation systems all make use of stochastic motion simulators. As well as the existing works that use particle systems for visualisation in performance contexts (described in section 2.5.4), there is theoretical grounding for their use as tools for timbre visualisation. In Smalley's (1994) discussion of spectromorphology and the idea of spectromorphological visualisation, it is highlighted that the perception of timbre evolution over time can be very dependent on the musical context. This is supported by research from experimental psychology. Smalley also mentions that a time-line view of timbral temporal variation can 'freeze the experience of temporal flux.' Stochastic motion simulators are capable of producing widely varying behaviours and textures. Their parameters can also be tweaked such that they leave more or less 'traces' of previous events. For example in fluid simulations this can be achieved by varying the dissipation values, and in particle simulations by varying the particle lifetime. Thus, it was decided that stochastic motion simulators could be good tools for the real-time visualisation of timbre and temporal spectral variation during live performances. The ability of stochastic motion simulators to produce widely varied motion and morphology supported the aim of Analema Group to explore relationships between sound, colour, light, movement and form.

4.4.3 Incloodu 2016

The 2016 'Incloodu Deaf Arts Festival' was held in London on 27th February. It involved artworks and installations produced by and for the international deaf community. Analema Group produced a performance and interactive workshop that involved real-time visualisation of audio during a live performance and generation of haptic feedback from the audio data that was experienced by audience members through Subpac technology.

Figure 4.12 shows a still image of the performance. A short video from the performance can be viewed at the following link:

https://www.youtube.com/watch?v=f9O8KIIYsmc

Performance Description

The performance was centred around two instruments with distinct sound characteristics and timbres: Tibetan singing bowls and a Japanese Taiko drum. These instruments were accompanied by a synthesised drone sound that continued throughout the performance. The performance followed a general structure involving sections where each instrumentalist improvised alone and sections where they improvised together. As a performance aimed at a hearing impaired audience, one of the main objectives was to represent the opposing characteristics of the two main instruments visually.

The visualisation was produced using a fluid simulation algorithm. Real-time audio data was

Figure 4.12: A still image from the performance at Incloodu 2016

captured from each instrument individually and analysed separately. The extracted audio features from each instrument were used to influence the velocity, density and colour values of separate emission points in the fluid simulation. The fluid simulation was thus used to artistically represent the behaviour and characteristics of the sound of each of the two main instruments, and the differences between them. Figure 4.13 shows an illustration of the technical performance setup.

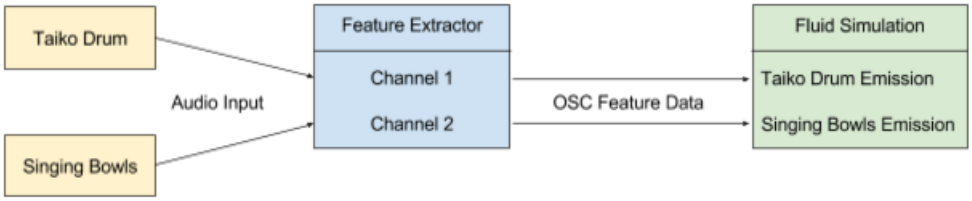


Figure 4.13: Incloodu 2016 - technical performance setup.

Visualisation

Onsets from the Taiko drum would trigger density and velocity to be entered from the top and bottom of the fluid grid. This produced individual streams of fluid falling from the top and rising

from the bottom of the fluid grid. The audio signal from the singing bowls was used to control the velocity, density and colour values of fluid in the centre of the grid. This created a central emission point from which fluid radiated outwards in order to complement the vertical fluid movement produced by the Taiko drum.

Mappings

The root-mean-squared amplitude of the audio from the Taiko drum controlled the velocity of the fluid at the top and bottom of the grid. This produced behaviour in the movement of the fluid that emulated the amplitude envelope of the drum hits; streams of initially high-velocity fluid that quickly decayed in momentum as the amplitude decayed.

The spectral centroid of the audio from the singing bowls controlled the brightness of the colour in the fluid at the centre of the grid. The root-mean-square amplitude of the singing bowl's signal also controlled the velocity at the centre of the grid. Again, the amplitude envelope of the singing bowls was emulated in the movement of the fluid. The movement was more constant than that of the Taiko drum, starting out with less rapid velocity increase and fading out over a longer period. The differences between each bowl were also represented, as 'duller' bowls produced less brightness in the fluid.

Discussion & Review

The visualisation and mappings used during the Incloodu performance were inspired by spectromorphological concepts. As mentioned in section 1.1.4, perceptually meaningful real-time musical timbre visualisation is dependent on the overall timbre make-up (the timbre space) of the performance. In the Incloodu performance one of the main focuses was the interaction and contrast between the timbre of the Taiko drum and that of the singing bowls. The visualisation of the Taiko drums covered a large area of the fluid field whereas the singing bowls visualisation came from a central emission point that gradually radiated outwards. This was intended to highlight the differences in spectral contents of the different instruments; the Taiko drum onsets had broadband spectral energy density, whereas the singing bowls had (relatively) narrow-band spectral content. In spectromorphological terms, the Taiko drum onsets sometimes 'masked' the continuously resonating singing bowls. This was also represented in the visualisation as the fluid from the drum would traverse towards the centre of the grid and interact with the fluid being affected by the bowls.

The visualisation during the Incloodu performance played a mainly artistic role. The layout of the emission points in the fluid simulation were designed in order to highlight the spectromorphological motions and behaviours of the two main instruments, and the differences between them. Although there were spectromorphological concepts that influenced the development of the visualisation system, the mappings used were not very complex. The main audio feature driving the visualisation was amplitude, with the addition of spectral centroid for the brightness in the case of the singing bowls. One key aim of future performances therefore become the use of spectral features as a key driving force for the visualisations, in order to achieve a more direct perceptual link between aural and visual stimuli. In the case of the Incloodu performance, the link was more conceptual.

4.4.4 Baltic Art Form 2016

The second Baltic Art Form Festival was held in London from the 3rd to the 5th of June 2016. The festival hosted performances and works from a collection of artists from the Baltic region along with British collaborators.

Performance Description

An audio-visual performance was presented by Analema group in collaboration with a choir. The performance took place in Saint Sepulchre's Church. The performed work consisted of a multilingual poem and accompanying composition for voice. The choir performed the composition continuously during which, at certain points, audience members were invited to the stage to read out specific lines of the poem. Lines read by the audience members cued specific sections of the choir performance. The work aimed to explore the relationships between music and light, sound and colour, spoken word and written language.

The audio signals from the choir and the audience microphones were captured separately in real-time and analysed individually using the Feature-Extractor application. The visualisation was projected directly onto the back four inner pillars of the church such that the architecture came to life with visual representations of the aural interactions happening during the performance.

Visualisation

The visualisation was produced using the Unity game engine and the built-in particle system editor that it provides. Unity also provides 'Wind Zone' objects that can be used to manipulate the movement of the particles. These were used to vary the movement of the particles from flowing and regular to rapid and chaotic. The generated visuals were duplicated and projected from two projectors, each projecting onto two of the back four pillars of the church. For each pillar, two particle systems were used for visualisation - one to represent the audio from the choir and one to represent the audio from the audience members. The technical performance setup for the Baltic Art Form performance is shown in figure 4.14.

Figure 4.14: Baltic Art Form 2016 - technical performance setup.

Initially the choir particle system was situated at the bottom of the scene and particles flowed upwards, with audience input producing particles from the top that flowed downwards and interacted with those of the choir. Throughout the performance, the emitters moved towards the centre

of the pillars and crossed paths such that, by the end, they had swapped places.

Mappings

The mappings used between timbre features and particle system properties were the same for both the choir and the audience audio signals. This ensured that differences between the timbre of the choir and that of the audience-member performers were directly represented in the visualisation. The emission rate of the particle systems was controlled by the root-mean-square amplitude of the audio, such that the particle streams were more pronounced for louder audio. As mentioned previously in section 4.4.3, one of the key objectives of this performance was to explore a direct perceptual link between the spectral distribution of the audio and the resulting visualisation. The spectral centroid of the audio was mapped to the strength of the external forces (Unity 'Wind Zone' objects) that manipulated the movement of the particles. Higher spectral centroid caused the movement to become more 'active' and chaotic. This was effective in visualising the difference between pitched audio (such as singing vowels) and audio with more varied spectral content (such as fricative sounds and whispering). A link between activity and spectral centroid was identified in Alluri and Toiviainen (2010).

Discussion & Review

In spectromorphological terms, one of the main compositional devices put to use in the performance was the use of contrasting sections of 'standard' vocal gestures (such as sung vowel sounds) and more abstract vocal textures involving variation from note gestures to noise textures. As mentioned in the previous subsection ("mappings") this was highlighted in the visualisation through the use of the spectral centroid to control the rate of motion in the particles. These vocal gestures and textures were often used in a 'streaming' behaviour, where some members of the choir would sing continuous tonal vowel sounds while others produced noisy, chaotic, sometimes percussive tones. In the visualisation, the continuous tonal vocal gestures produced continuous streams of particles on the columns. The noisy fricative vocal textures would then cause the particle stream to rapidly diffuse in chaotic movement. Thus, the interaction between the different textural streams in the spectrum was visualised in the behaviour of the particles.

Although the mapping from spectral centroid to force strength produced a direct link between the spectral distribution and the behaviour of the particles, the mapping strategy was still fairly limited in scope. A key objective for the next performance was therefore to explore many different mapping strategies and different ways in which timbre can be represented by making use of different perceptual analogies.

4.5 Ron Arad's Curtain Call 2016

The previous section gave descriptions of two pilot-study performances that were carried out in order to explore the use of perceptual timbre representation in an artistic context, to influence the design of real-time visualisations during audio-visual performances. This section will go into

detail about a third case study performance for which evaluation data was collected from both audience members and performers. A description of the event is given, along with descriptions of the visualisation system and the development of the mapping strategies that were used during the performance. This is followed by a presentation of the analysis that was carried out for the performance. Finally, the performance and the development process involved in the performance are discussed and reviewed.

The Roundhouse in London hosted Ron Arad's Curtain Call installation from 6th - 29th August 2016. Ron Arad's Curtain installation makes use of 5600 thin cylindrical silicon rods suspended from a circular ring 18 metres in diameter. This provides a 360 degree canvas onto which videos and animations are projected using 12 projectors arranged in a circle around the ring. The Roundhouse also hosted a number of immersive live performances designed around the 360 degree projection technology. One such performance was produced by Analema Group. It was focused on the representation of vocal qualities in contrasting ways. The event featured numerous sections varying between prearranged performance by a group of artists and open audience interaction where microphones were open for audience members to make sounds and observe their real-time representation on the 360 degree projection.

Figures 4.15 and 4.16 shows some still images from the event. A short video of the performance and installation can be viewed at the following link:

https://www.youtube.com/watch?v=yGBPjv2Sgbk

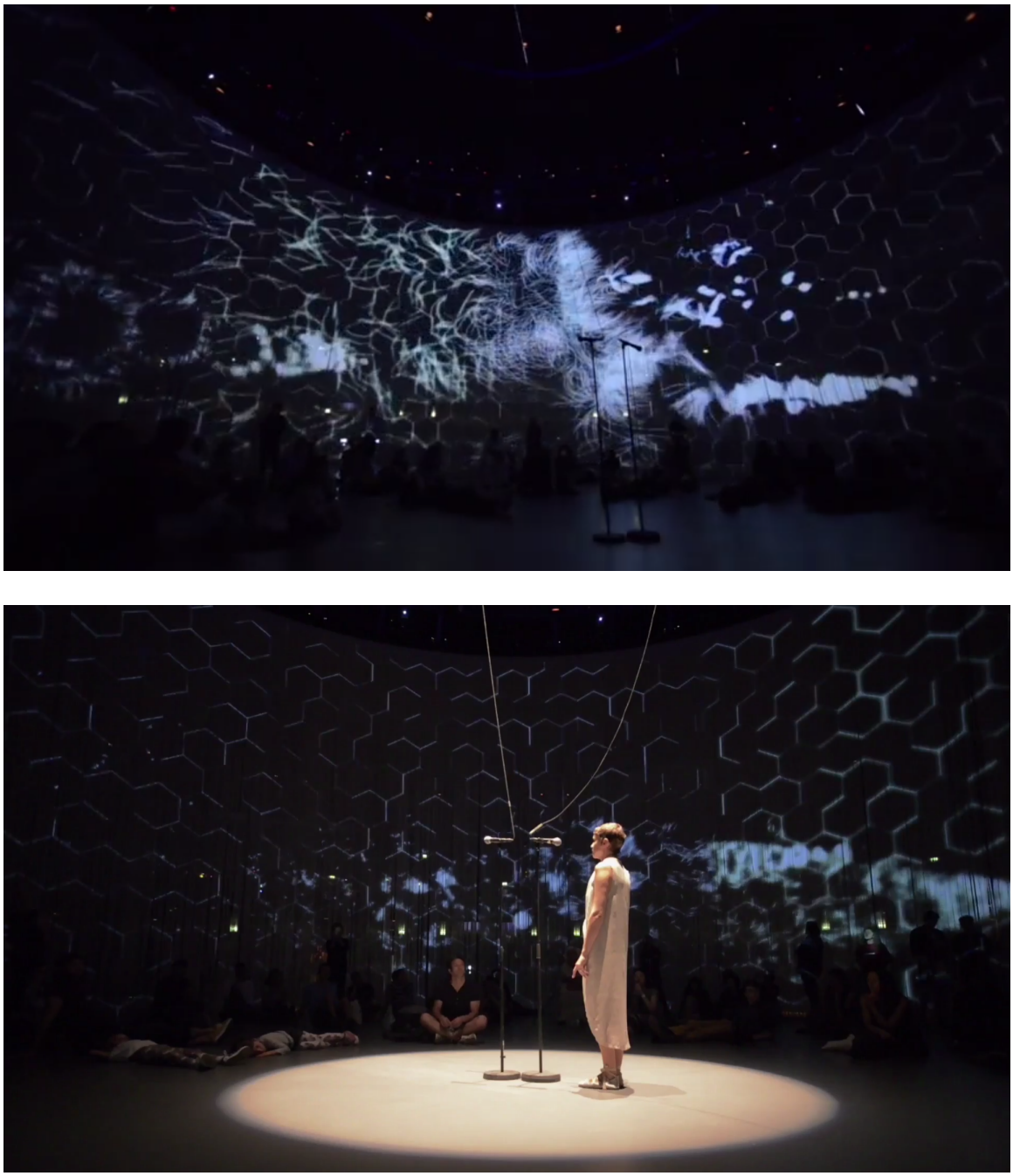


Figure 4.15: Still images from inside the curtain installation during the performance at Curtain Call 2016

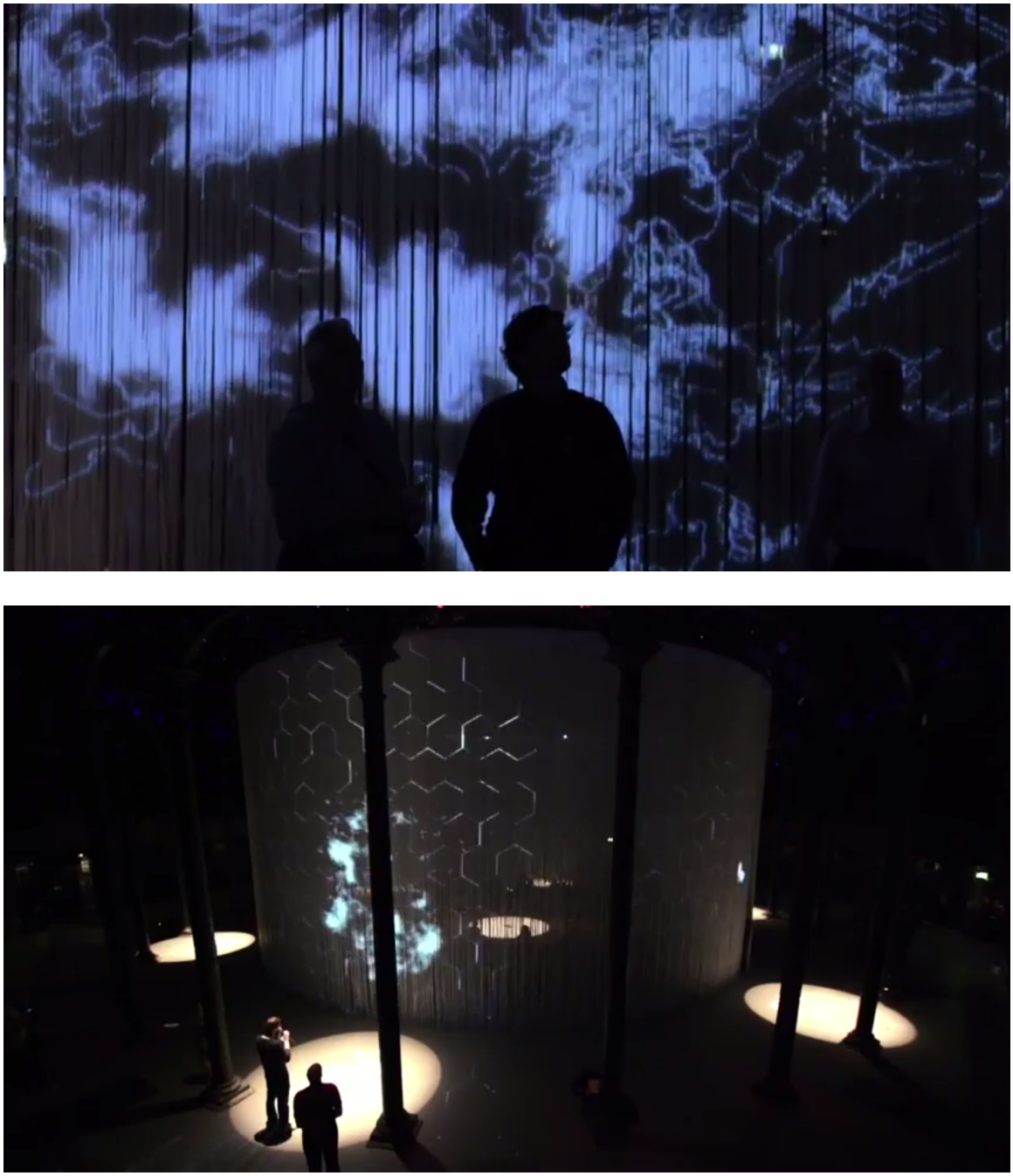


Figure 4.16: Still images from outside the curtain installation during the performance at Curtain Call 2016

4.5.1 Performance Description

The musical content of the performance focused on contrasting vocal qualities, specifically the contrast between tonality and noise. The main aim of the performance was to present contrasting representations of the difference between pitched sound and noise, by visualising the sound from different microphones using different mapping strategies and visual aesthetics. The artistic motivation came from the attribution of meaning to sound, and the ways in which visual analogies might be used to describe abstract sound qualities. While the previously described performances made use of specifically designed low-level audio-visual mappings for timbre representation, this performance explored the concurrent use of *contrasting* mapping strategies in order to demonstrate how similar sound attributes can be represented in contrasting ways by drawing attention to specific audio-visual perceptual correlations. To this end, neural networks were used such that mapping strategies could be designed on a 'higher level' with reference to general visual aesthetics and through the direct use of vocalised audio input data. This process will be described in more detail later in this section.

During the performance, six microphones were arranged around the outside of the curtain and there were two microphones in the centre of the curtain. The Feature-Extractor application was used to analyse the audio signal from each microphone. A custom particle system was developed and the audio features controlled parameters of the particle system. Each microphone around the outside of the curtain controlled a separate emission point in the visual particle system, and produced unique visual characteristics in response to varying audio qualities. The microphones in the centre were used to control the behaviour and visibility of an underlying grid structure that defined global characteristics of the movement of the particles, across each visual world.

Visualisation

The visualisation system was developed using the Unity engine by Dario Villanueva, a member of Analema Group. The system consists of a number of particle emitters and an underlying grid of forces. For each emitter, the effect of these forces on the particle movement can be parametrically controlled. The direction of these forces is gradually perturbed according to a Perlin noise function, the parameters of which can also be parametrically controlled. When particles are generated or 'birthed', they are birthed with a given amount of life. This life value is then decreased at a certain rate every frame, and the particle dies and new particles are birthed from the emitter. The various parameters of control for the particle system are:

- **Current Life:** The level of life the particles are generated with.
- **Life Decay:** The rate at which at particles decay and die.
- **Bloom:** A post-rendering glow effect.
- **Hue:** The colour tint of the glow effect.
- **Force:** The magnitude of the external forces.
- **Height Offset:** Vertical movement of the particle emission point.

- **Particle Trail:** The length of trails left by the particles. Small values produce individual particles. Large values produce long linear trails.
- **Max Speed:** The maximum speed of the particles.
- **Noise Time Offset:** The magnitude of temporal change in the Perlin noise function. Smaller values produce similar values from frame to frame (gradual change in direction). Larger values produce more distant values from frame to frame (staggered jittery change in direction).
- **Spread Factor:** The extent to which particles are bunched together or spread out. Small values produce emission in bursts. Larger values produce continuous emission.

As mentioned previously, there were six particle emission points used in the Roundhouse performance corresponding to each of the six microphones around the outside of the curtain. Audio characteristics from each individual microphone were mapped to the different particle system parameters in different ways, and using different ranges, in order to produce contrasting visual characteristics for each individual region on the curtain. A controller app was developed that can be used to control the ranges of each of these parameters via OSC messages sent to the visualisation system. The controller app can be used to save and load presets and to move the emission points around the visual space. This was important when creating and arranging the contrasting visual regions in the visualisation space for the Roundhouse performance.

There were various networked machines running individual software applications during the performance. One machine was used to add sound design and real-time sound editing to the choir's performance. Another machine was dedicated to real-time audio analysis. Both of these machines received audio input from the front-of-house audio desk. Two other machines were used to run the neural networks and controller app, respectively. Finally two 7thSense Infinity Delta Media servers were used to run two instances of the Unity visualisation app, which was projected onto the curtain. All of these machines and applications communicated via OSC through an Ethernet switch, as shown in Figure 4.17.

Mapping via Machine Learning

The Wekinator software, developed by Rebecca Fiebrink, facilitates the use of on-the-fly machine learning for the development of interactive performative controller applications (Fiebrink, Trueman, and P. R. Cook 2009). The Wekinator software was used for the Roundhouse performance in order to train individual neural networks for each of the six microphones around the outside of the curtain. The features extracted by the Feature-Extractor for each of the microphones were used as inputs and neural networks were trained to control individual parameters in the particle system (via OSC) for each of the emitters in response to specific combinations of audio features. The training process involved first setting the OSC control sliders in the Wekinator software in order to produce the desired visual response. Sound input was then produced and fed through the Feature-Extractor, which output features via OSC to the Wekinator input. This recorded audio feature data was used to train Wekinator to associate the given output values (as set on the Wekinator GUI) with the audio feature data. Subsequently, the Wekinator GUI was used to produce a contrasting visual output in the particle system, which was then associated with contrasting audio characteristics using different training data. Once the networks had been trained with varying audio characteristics, they could be run in real time, and new incoming audio feature data from

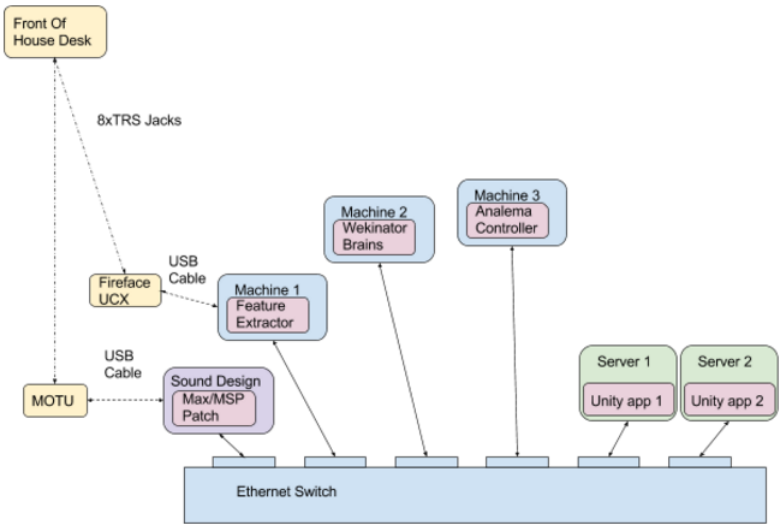


Figure 4.17: Curtain Call 2016 - technical performance setup.

the Feature-Extractor would be referenced against the existing trained models in order to produce desired visual output. During the final performance, there were seven neural networks running in real time: Six for the individual microphones around the perimeter of the curtain and one for the microphones in the centre of the curtain.

The controller network for the central microphones was trained to control the parameters of the Perlin noise function that perturbed the directions of the external forces. The central microphones were used as very simple controllers, and as such only the amplitude was mapped to visual parameters. The X- and Y- resolution of the Perlin noise function were both altered when sound was made at the central microphones such that the overall behaviour of all of the particles across all visual regions was altered in response. In the visualisation system, the external forces can be rendered as directional vectors which form a hexagonal grid. The lines of this underlying grid were rendered with a certain opacity level. The amplitude at the central microphones was mapped to this opacity level such that making sound would highlight the underlying grid that was influencing the movement of the particles in the system.

For the six individual visual regions, the main audio characteristic that was visualised was the level of periodicity versus noise. During training of each of the networks, tonal and noisy sounds were vocalised in order to produce real-time training data on-the-fly. In each region, different visual parameters were used to represent the difference between noise and tonality in different ways. For example in one of the regions, the controller network was trained to associate tonal sounds with a higher bloom value, such that tonal sounds produced a glow effect and noisier sounds produced more fine-grain texture by showing individual particles. In another region, noisier sounds were associated with greater particle movement speed and lower external force effect, such that a noisy signal would produce fast radial particle emission in comparison to slow, directed particle emission from tonal signals. In contrast, a different region associated noisier signal with greater external force effect and greater speed. This created quick angular movement in the particles in response

to noisier signal, in contrast with slower curved movement in response to tonal signal.

4.5.2 Evaluation

Two surveys were conducted in order to evaluate different aspects of the visualisation system in the context of real-time performance. The first survey was conducted with audience members and the second with performers. The questionnaires in each survey made use of 7-point Likert-scale ratings, from "Strongly Disagree" to "Strongly Agree". Each survey was implemented and conducted online using Survey Planet. In the following sections the questionnaires will be described and the results will be presented, along with analysis and evaluation.

Likert Scale Data

There is a debate around the use of parametric analysis techniques (e.g. t tests or analysis of variance) when analysing Likert scale data (Carifio and Perla 2008). Of particular importance in this regard is the difference between a Likert *scale* and a Likert *item*. In the context of a questionnaire, a Likert item consists of one single question. A Likert scale consists of multiple questions that can be *combined* to measure some kind of general attitude or opinion. since the data obtained from a single Likert *item* is fundamentally ordinal in nature, it does not make statistical sense to carry out parametric statistical analysis on such data. However, when multiple Likert items are combined to form a Likert scale, the resulting data is interval in nature (Norman 2010). It is important to note that the data is interval in nature because the comparison is between number of responses between questions, for each Likert option (i.e. not between individual Likert options). It is therefore only advisable to use parametric analysis on Likert scales when the number of questions that form the scale is sufficiently high. In the questionnaires presented below, the number of questions in each Likert scale is low, and each scale contains a negatively phrased question. The median and interquartile range (IQR) are therefore used when reporting and analysing the results. The individual items (questions) in each Likert scale are compared using their median and IQR values, and some potential implications are drawn.

Audience Questionnaire

The key aim of the audience questionnaire was to measure how effective the visualisation system was in enhancing and supporting the overall performance, and how meaningful the visualisations were in the context of the performance. In this context, 'meaningful' refers to the levels of connection and association that were perceived by the audience members between visual and aural stimuli. The sonic aspects of the performance focused on contrasting timbral qualities (between tonal sound and noise) as opposed to structured pitches and rhythms. Reports of perceived connections between aural and visual stimuli during the performance would thus indicate that the visualisations were effective examples of perceptually motivated timbre visualisation.

There were 14 participants in total for the audience questionnaire (7 male, 7 female). The full set of questions and their results are provided in appendix A.1. The questions were grouped according to their objectives as follows:

- Questions 1 - 3: Demographic information
- Questions 4 - 7: Existence of connection between aural and visual stimuli.
- Question 8: Level of importance of aural stimuli
- Questions 9 - 12: Understanding of connection between aural and visual stimuli.
- Questions 13 - 16: Ability to intentionally control visual response through aural input.
- Questions 17 - 22: Feedback specific to Analema Group.

In the questionnaire these objectives weren't explicitly stated; all questions were presented sequentially as one group. Questions 17 - 22 were included for Analema Group as useful feedback for their future performances. The analysis below therefore focuses on questions 4 - 16.

Results

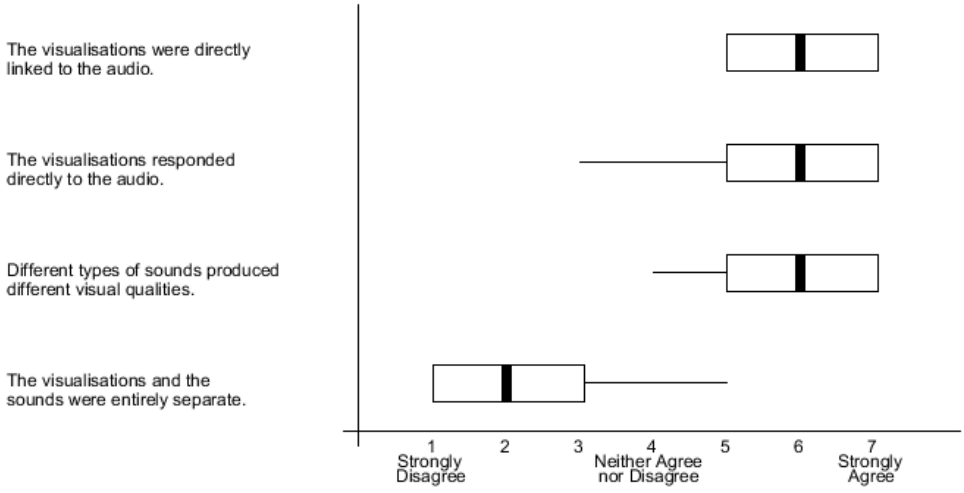


Figure 4.18: Audience Questionnaire Questions 4 - 7.

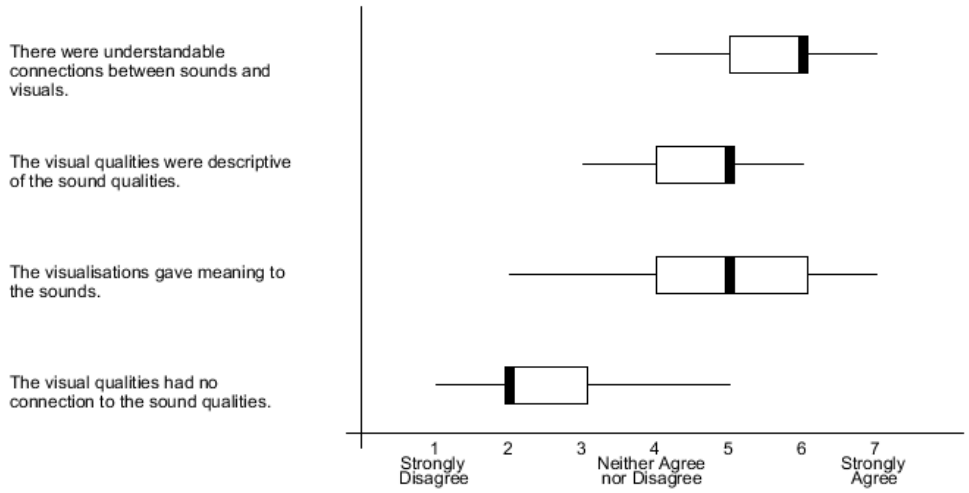


Figure 4.19: Audience Questionnaire Questions 9 - 12.

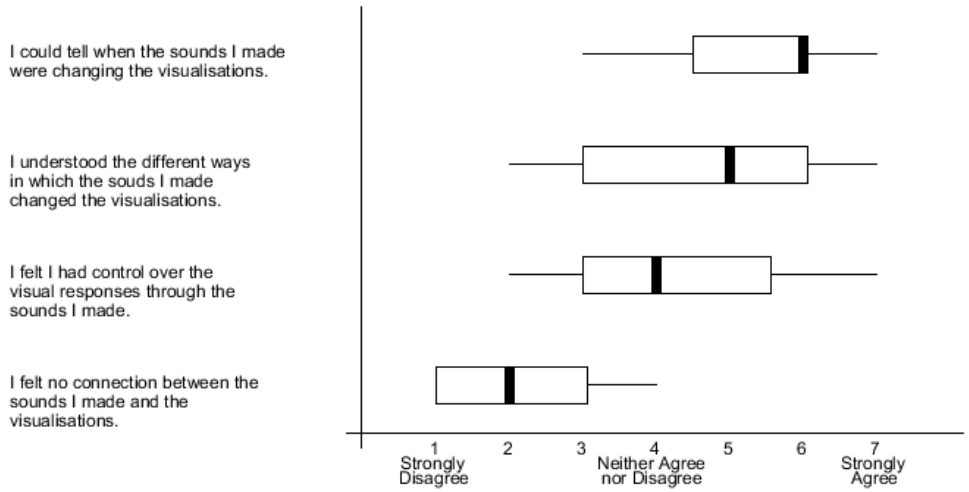


Figure 4.20: Audience Questionnaire Questions 13 - 16.

Analysis

As shown in Figure 4.18, there was general agreement that the sounds and the visuals were connected during the performance. The IQRs span from "agree" to "strongly agree" for the positively phrased questions and from "strongly disagree" to "disagree" for the negatively phrased question.

This suggests that there was agreement as to the presence of a connection between sounds and visuals, but that there was some disagreement as to the strength of this connection. There was slightly more uncertainty as to the comprehension of the connection, as shown in Figure 4.19. For question 9, the IQR suggests agreement that the connections were understandable on some level. Question 12 suggests agreement as to the presence of a connection between individual visual qualities and individual sound qualities. Questions 10 and 11 suggest less strength of agreement. This could be due to the fact that the questions were more subjectively worded, and asked participants to grade the descriptiveness and meaningfulness of the connections. For such subjective questions, it would be beneficial to examine a much larger population to obtain more generalisable results. Again for questions 13 - 16 (Figure 4.20), questions 13 and 16 suggest agreement as to the perception of a connection between sound and visuals when interacting with the system. Questions 14 and 15 suggest more uncertainty as to the feeling of directed control over this connection.

Performers Questionnaire

The key aim of the performers questionnaire was to measure the level of interaction between the artists and the visualisation system. The questions were designed in order to examine the level of control the artists felt over the visualisations, and the extent to which the artists perceived the visual representations as important factors within their performances. In this way, the questionnaire examines how effective the perceptual timbre representation system was as a performance *tool*.

Each artist involved in the performance completed the survey, therefore there were six participants. The full set of questions and their results are provided in appendix **??**. The questions were grouped according to their objectives as follows:

- Questions 1: Previous experience
- Questions 2 - 5: Existence of connection between aural and visual stimuli.
- Questions 6 - 9: Ability to intentionally control visual response through aural input.
- Questions 10 - 13: Importance of visual response as a factor in the performance.
- Question 14: Interest in future collaborations.

In the questionnaire these objectives weren't explicitly stated; all questions were presented sequentially as one group. The analysis below focuses on questions 2 - 13, since questions 1 and 14 were more specific to Analema Group collaborations.

Results

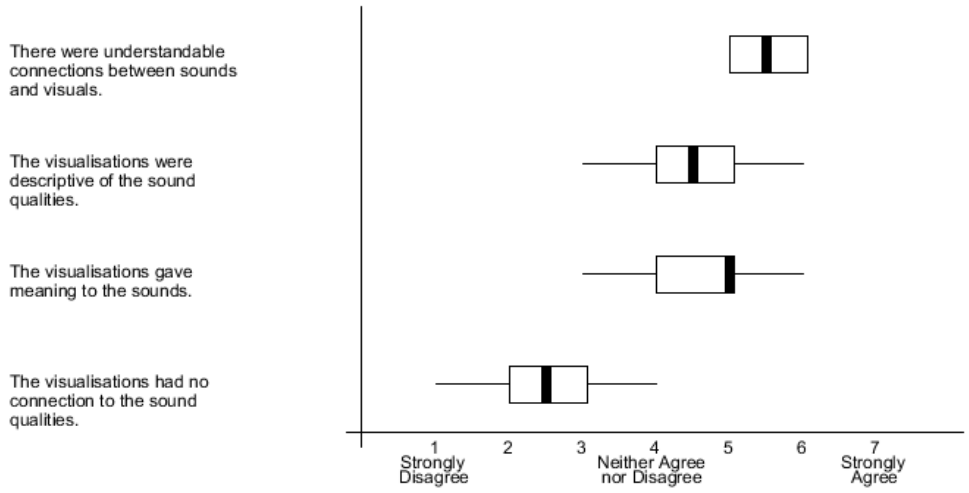


Figure 4.21: Performers Questionnaire Questions 2 - 5.

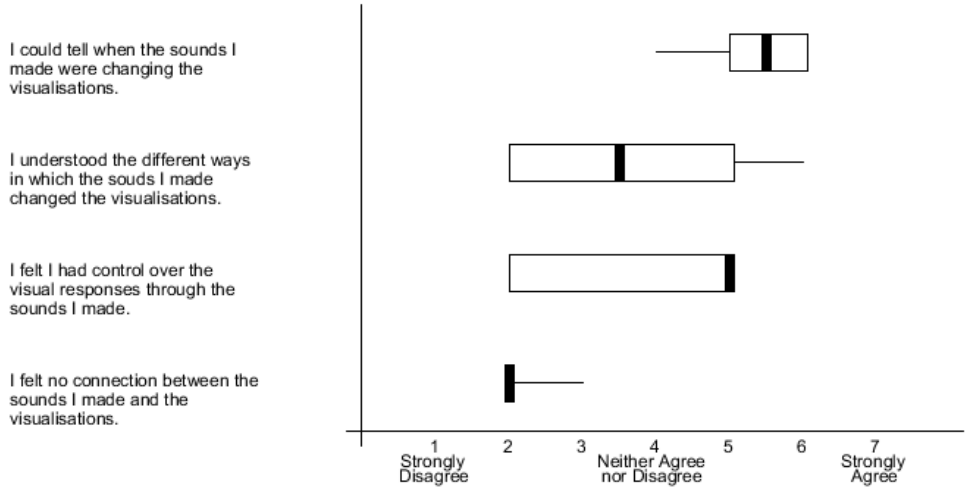


Figure 4.22: Performers Questionnaire Questions 6 - 9.

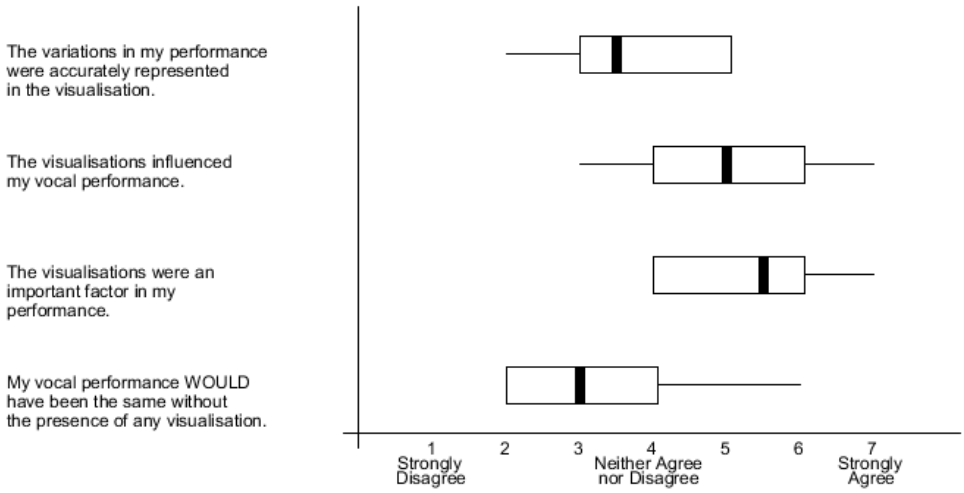


Figure 4.23: Performers Questionnaire Questions 10 - 13.

Analysis

Figure 4.21 suggests the performers perceived a connection between their sounds and the visuals. Questions 3 and 4 from the performers questionnaire suggest less strength of agreement as to the meaning and descriptiveness of the visualisations. This is a similar outcome to that of questions 10 and 11 from the audience questionnaire (Figure 4.19). Figure 4.22 suggests that the performers agreed that their vocal input had an effect on the visualisations. However, there was more uncertainty as to the *level* of control that they experienced. As mentioned previously, the visualisation system was separated into six independent zones for each performer, which had varied mapping strategies resulting from different training models. The variation in reported level of control could be due to this variation in mapping strategies. Figure 4.23 suggests that the performers felt the visualisations influenced their performances. However, the results of question 10 show uncertainty as to the accuracy with which the vocal variations were represented. Again, this is likely due to the variations in mapping strategies across the six zones.

4.5.3 Discussion & Review

The Roundhouse performance was designed around the use of contrasting audio-visual mapping strategies to show how abstract sounds can be represented in different ways, by drawing attention to specific audio-visual perceptual correlations. Six performers provided real-time musical input ranging from tonal and harmonic singing to abstract, atonal and percussive vocalisations. Each performer was associated with an individual visual region in which their sound input was uniquely represented, contrasting with neighbouring representations of similar (and at times contrasting) sound input. The performance was intended to demonstrate the numerous possibilities available when representing sound visually, and to draw attention to the subjectivity involved in the meaning

of a sound, by providing simultaneous contrasting visual representations (and suggested meanings or interpretations) of similar sound qualities. Different correlations between specific semantic qualities and specific acoustic qualities of timbre were used to produce contrasting visual representation methods.

The use of machine learning, and in particular the use of the Wekinator software and the technique of on-the-fly training was very appropriate in the development of the mapping strategies used in the Roundhouse performance. On-the-fly training of a neural network facilitated the development of mapping strategies from a categorical (rather than parametric) perspective. Such training allows the user to think in general terms. Specific visual qualities can be designed and setup, and then can be easily associated with specific audio characteristics through direct sound input and fast training. This is particularly advantageous when working with artists and creators without technical knowledge of sound analysis, machine learning, or generative visualisation. By working directly with desired visual output and real-time example sound input, arbitrarily complex mapping strategies can be quickly implemented and evaluated. As such, this kind of development was very much in line with the motivations of the performance, which aimed to explore how specific correlations between different semantic and acoustic qualities of timbre can be used to produce contrasting timbre representation implementations.

Although the Wekinator software and the technique of on-the-fly training was put to great use in the development of the Roundhouse performance, there is large scope for further development in this area. For the Roundhouse performance, the Wekinator controller networks were trained by the development team and one of the performers. The training data then generalised to the individual performer that was assigned to that specific visual region for the actual performance. However, the Wekinator software could actually be used to distinguish between individual performers, and adjust the visual output accordingly. This would provide a way in which specific visual aesthetics could be associated with specific voice types, using a single microphone. As mentioned previously, the Roundhouse performance focused mainly on the difference between noisy and pitched vocalisations. However, the combination of the Feature-Extractor and the Wekinator machine learning tool could be used in the recognition of more complex aural structures. For example, a controller network could be trained to recognise the difference between vowel types, the difference between voice and other sounds, the differences between individual voice types, or any other axes of sound discrimination. Going beyond vocal sounds, there is massive scope for further research in the use of on-the-fly training of machine learning algorithms to develop complex representation strategies for specific families of timbre space.

Analysis of the visualisation system through questionnaires has revealed general agreement as to the presence of a real-time connection between audio and visuals during the performance. There was also agreement as to the descriptive quality of the visualisations, although this agreement was generally weaker. There was agreement that the visualisations could be varied by varying vocal input, but there was more uncertainty as to the level of this control, in both audience members and performers. The variation in mapping strategies that were produced for each zone likely contributes to the variation in reported levels of control over the system generally. This study was conducted in order to evaluate the use of a semantic timbre visualisation system 'in the real world' during a live performance. However, It would be useful to conduct further evaluations with larger numbers of participants in a more controlled environment.

4.6 TimbreFluid

The visualisations for the performances described in the previous section were developed in very specific contexts, with close reference to both the structure of the performances and the timbre spaces involved in the performances. As has been mentioned before, this is necessary due to the 'absolute' nature of the perception of temporal timbre variation. However, although the perception of timbral variation may vary depending on context, the use of semantic mappings for acoustic timbre features nonetheless still provides a good *general* starting point for the use of stochastic motion simulators for timbre visualisation. Once these mappings have been set up, their magnitude and weighting can be altered in order to suit the performance. This section will describe the development of a general real-time visualisation system for live musical performance based on a fluid simulation algorithm and mappings from acoustic timbre features to fluid simulation parameters.

The 'TimbreFluid' system aims to demonstrate a set of usable mappings from timbre features to common fluid simulation parameters that are based on correlation with semantic descriptors of timbre. In the case of Hsu (2011), similar timbre descriptors were used but the mappings were intentionally left 'open to interpretation'. TimbreFluid aims for the opposite: mappings that are intuitively understandable due to their grounding in correlation between semantic descriptors of timbre and acoustic features of timbre. This section will use the development of the TimbreFluid system and the specific mappings implemented in the system in order to highlight general mappings from acoustic timbre features to visual properties of stochastic motion simulation systems. This facilitates a specific extension of perceptual timbre representation based on specific visualisation techniques (stochastic motion simulators) and provides further demonstration of how perceptual timbre representation can be used in specific performance contexts.

An overview of the TimbreFluid system is given, followed by a description of the various visual parameters that can be altered in the fluid simulation algorithm. These parameters are broken up into groups relating to general visual properties. For each of these visual categories, more general discussions are given on how they can be parametrically controlled in other stochastic motion simulation algorithms. Following the descriptions of the visual parameters, a discussion is given on the various mappings that are used in the TimbreFluid system. Again, this is combined with general discussions on how mappings can be applied across different algorithms. Finally, the system is discussed and possible directions for future development are outlined.

4.6.1 Overview

The TimbreFluid demonstration system was developed in order to investigate different mappings for specific timbre features to fluid simulation parameters. As mentioned previously in section 2.5.4, stochastic motion simulators have predominantly been used in the context of congruent gesture-to-audio and gesture-to-visual mapping. In the limited number of works where stochastic motion simulators are used for direct audio visualisation, the audio features used are often limited (for example using only amplitude and pitch).

In the TimbreFluid visualisation, fluid is emitted in bursts from the centre. The properties of the fluid are then controlled parametrically by the audio features in real time. An example video of this system can be viewed at the following link: https://youtu.be/LnoJ6eLfAK4.

The Cinder C++ framework is used for the fluid simulation. The 'cinderfx' Cinder block (framework extension) is used, which provides a Fluid2D class. This class provides a number of fluid parameters which can be altered to provide various effects. These are detailed in this section.

4.6.2 Luminance / Brightness and Colour

The colour and luminance of the fluid is controlled using a hue-saturation-value (HSV) algorithm. The hue, saturation and value can be altered in real-time. Figure 4.24 shows the hue of the fluid varying from low to high. A colour distorting effect is implemented in the centre of the fluid. This effect produces increased colour contrast and dissonance in the centre of the fluid.

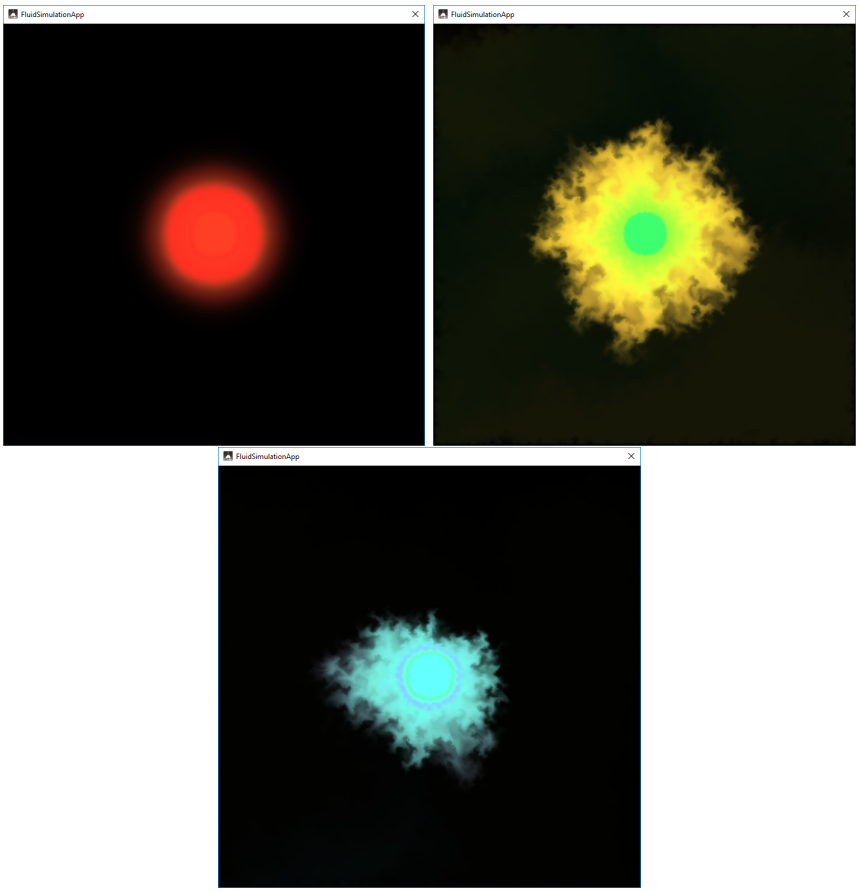


Figure 4.24: TimbreFluid - Varying hue from low (left) to high (right).

In general, luminance, brightness and colour properties can be controlled parametrically by mapping audio features to RGB or HSV values. Similarly, audio features can be used to drive procedural texturing algorithms and rendering algorithms.

In fluid simulation algorithms, there is usually a colour grid where RGB or HSV values of

individual grid positions can be altered parametrically. In particle simulation algorithms, particle colours and textures can be controlled parametrically.

4.6.3 Texture

The parameters of the fluid simulation that are altered in order to vary the texture are: RGB viscosity, velocity viscosity, RGB dissipation, velocity dissipation, and vorticity scale. Figures 4.25, 4.26 and 4.27 show the effect of varying these parameters.

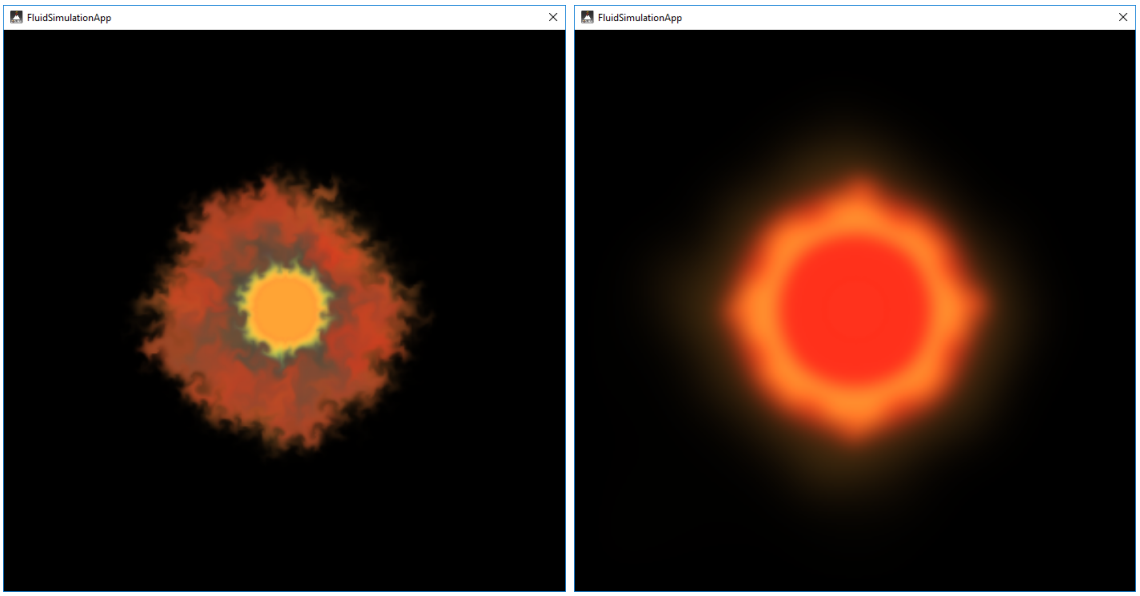


Figure 4.25: TimbreFluid - Varying viscosity from low (left) to high (right).

In fluid simulation algorithms the 'viscosity' parameter models the fluid's gradual loss of kinetic energy due to friction (Müller, Charypar, and Gross 2003). Higher viscosity values will therefore lead to slower, less chaotic movement. A lower viscosity value leads to faster, less resistive fluid movement.

The dissipation value affects the rate at which the energy transfer throughout the system decreases. As explained by Stam (2003), dissipation is necessary in fluid simulations in order to keep the algorithms stable. However, numerical dissipation can lead to the simulations dampening *too* quickly (Stam 2003). 'Vorticity confinement' is a technique whereby energy can be re-introduced back into the system and produce vorticity (swirling movement) at local scales (Fedkiw, Stam, and Jensen 2001). The 'vorticity scale' parameter in the cinderfx fluid simulation affects the scale at which this vorticity is produced. Lower vorticity scale leads to fine-grain textures and high vorticity scale produces coarser textures in the fluid.

In general, complex textures are produced by parametrically controlling the movement and behaviour of the fluid or particles in a stochastic motion simulation algorithm. As discussed, in fluid simulations this can be achieved by altering the vorticity amount, vorticity scale and

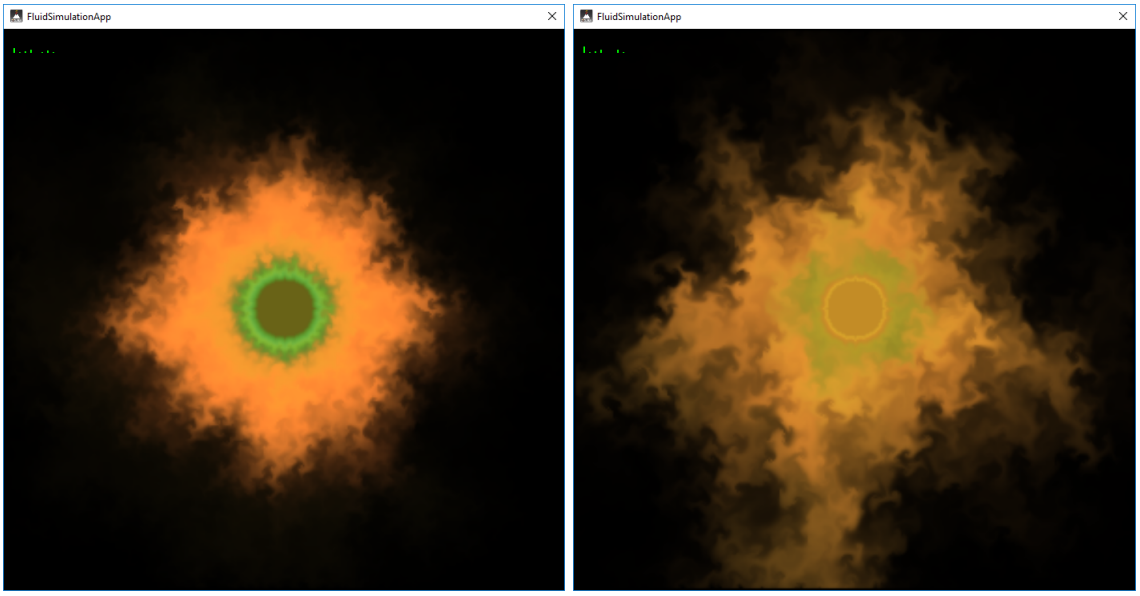


Figure 4.26: TimbreFluid - Varying velocity dissipation from low (left) to high (right).

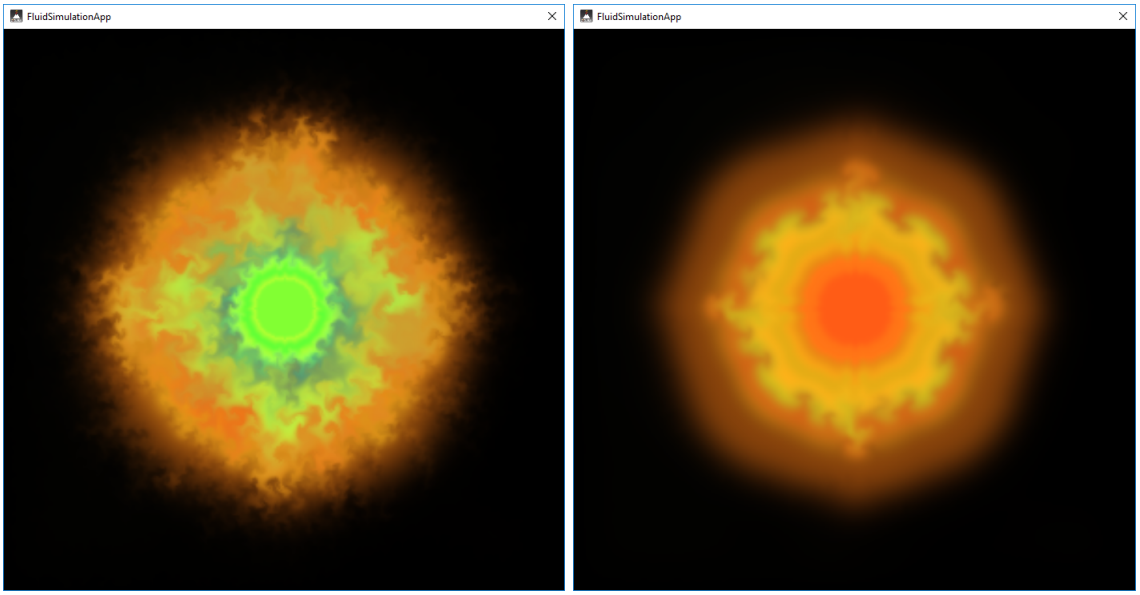


Figure 4.27: TimbreFluid - Varying vorticity scale from low (left) to high (right).

dissipation rates. Fluid simulations also often include control parameters for force modifiers such as gravity. In particle systems, properties of the environment can be altered in order to control the movement of the particles. For example, the point of emission (where particles appear and begin their movement) has numerous properties. These include the emission rate and the shape (e.g. sphere, cone, plane). Similarly, the particles can be affected by external forces in order to produce chaotic movement. For example in the Unity game engine, 'wind zones' can be created that apply directional force to particle streams. Particle systems can also make use of collision areas, which cause the particles to collide and bounce, altering their trajectory. Direct parameters of the particles can also be controlled such as movement speed and length of lifetime.

4.6.4 Mass / Volume

In TimbreFluid the fluid is emitted from a central emitter. The bursts are emitted with a certain velocity (force) which affects how far the fluid initially travels. This can be controlled parametrically. The RGB dissipation values can also be used to control how much fluid remains active in the system over time. Lower dissipation rates will lead to fluid staying active for longer, as more energy is introduced to the system. This makes the overall fluid system 'fuller' as it becomes filled with active fluid that doesn't quickly dissipate away.

In stochastic motion simulation algorithms, density and mass can be controlled directly by varying the density of the fluid or particle stream. In fluid simulations the density grid and density dissipation can be used to control the density of the fluid. The viscosity of the fluid can also be altered such that the behaviour and motion of the fluid resembles that of a more – or less – dense fluid. In particle systems, the emission rate and particle count can be used to alter the density of the particle stream. Texturing and colour effects can also be used to alter the apparent density of individual particles. External forces and particle sensitivity to these forces can also be used to simulate the movement of dense or light materials or substances.

4.6.5 Mappings

Luminance / Brightness

In the construction of the HSV colour for the continuous bursts of fluid, the following mappings are implemented:

- Hue - spectral slope
- Saturation - spectral spread (inverse)
- Value - (spectral centroid + RMS) * RMS

The spectral slope is used to sample a colour from the hue wheel. Spectral slope is linearly dependent on spectral centroid, so this mapping has a similar effect to the centroid-hue value mapping described in the Sound Signature tool. Audio with more low-frequency content will produce lower hue values, and vice-versa. The spectral spread is inversely mapped to the colour

saturation. Therefore, narrowband signals will produce very saturated colour, which has been sampled from the hue wheel according to the spectral spread. Conversely, broadband signals will produce more diluted colour. The spectral centroid controls the value (brightness). As mentioned previously in section 1.5, there are various existing studies that suggest spectral centroid as a good indication of the perceived 'brightness' of a sound. The colour distortion effect in the centre of the fluid is used to represent the harmonic features. The amount of brightness alteration is controlled by the inharmonicity and the resolution of the colour distortion is controlled inversely by the harmonic energy ratio.

As discussed throughout this thesis, there is strong correlation between spectral centroid and the perceived brightness of an audio signal. Similarly, the results from numerous studies show participant preferences for the mapping from spectral centroid to visual brightness. Real-time timbre visualisations using stochastic motion simulation algorithms can make use of this correlation by mapping the spectral centroid to the brightness of the fluid or particle stream. Frequency content has also been represented using colour in some existing systems. The survey described in section 3.3.3 shows a preference for the inverse mapping from spectral centroid to hue (thus mapping low frequencies to blues and high frequencies to reds). The systems described in this thesis have also made use of the inverse mapping from spectral spread to colour saturation. The Sound Signature survey showed a preference for this mapping over the absence of this mapping, in the context of waveform visualisations. This mapping can be implemented in stochastic motion simulations by implementing an HSV colouring algorithm and controlling the saturation inversely using the spectral spread value.

Texture

The various parameters of the fluid motion are parametrically controlled using the following mappings:

- Vorticity scale - spectral flatness
- Velocity dissipation - spectral centroid
- RGB dissipation - RMS amplitude
- Velocity viscosity - spectral spread (inverse)
- RGB viscosity - spectral spread (inverse)

The spectral flatness, which indicates how 'noisy' the signal is, controls the vorticity scale. This means noisier signals will produce more fine-grain textures in the fluid. As mentioned previously in section 1.5, this correlation has been reported in the existing literature. The spectral centroid controls the rate of velocity dissipation. Signals with higher spectral centroid values will produce more 'active' fluid. The spectral centroid gives an indication of the level of high frequency energy in the signal. By mapping spectral centroid to velocity dissipation, high energy in the signal will produce more rapid motion. Therefore, as well as semantic description, this mapping is based on the physical nature of sound. Higher frequencies follow from higher rates of change, and using this mapping they produce higher rates of change in the fluid behaviour.

As mentioned previously in section 4.4.2, textures can be produced and varied in stochastic motion simulation algorithms by varying the behaviour and motion of the fluid or particles. A physically-motivated mapping can be implemented by mapping high frequency content to more rapid rates of motion. This was implemented in the Baltic Art Form performance. The strength of different external forces was controlled by the spectral centroid. This produced fast chaotic movement in the particle stream when high frequency vocal techniques were used such as fricative sounds or whispering. Chaotic, multi-directional movement can also be used to create fine-grain textures. For example using vorticity confinement in fluid simulations, or using multiple combined directional forces to act on a particle system. The granularity of movement can be controlled by the spectral flatness in order to represent the difference between pitched audio (uniform movement) and noisy audio (multi-directional movement).

Mass / Volume

The initial velocity of the fluid or 'burst strength' is controlled by the root-mean-squared (RMS) amplitude of the signal. This way, the size of the initial motion is controlled by the volume. The RMS also controls the base brightness of the fluid burst (which is then augmented using the spectral centroid). An association between brightness and loudness was demonstrated in Giannakis (2006).

The low frequency energy ratio inversely controls the viscosity of the fluid. This means that signals with more low-end energy will produce more 'dense' fluid. An association between the 'density' of a sound and the low frequency energy content has been reported by Alluri and Toiviainen (2010).

4.6.6 Discussion

The fluid simulation visualisation system visualises the audio as a continuous system of fluid. The audio features drive the parameters of different parts of the system in real time. The fluid system is very complex, and the behaviour of the fluid can be quite unpredictable due to the global interaction of the fluid. For this reason, this kind of visualisation is more suited to solely the representation (rather than manipulation) of timbre. For example, generative visualisations such as this are often used in audio-visual performances and installations.

The stochastic nature of fluid simulations is often drawn upon in order to simulate real-world phenomena that also have seemingly stochastic properties (e.g. fire, water). One possible avenue of research could be to use the audio features of an audio sample of one such real-world phenomenon (or even a Foley artist's impression) in order to influence the graphic simulation.

4.7 Conclusion & Summary

The key objective of this chapter and of the systems described within has been to introduce and demonstrate perceptually motivated methodologies by which musical timbre can be visualised. Existing research into the perception and description of timbre suggests that it is often conceptualised

with reference to physical, material and visual qualities. The systems presented in this chapter attempt to draw upon perceptually and semantically meaningful mappings from low-level audio features to low-level visual features such that certain visual metaphors for timbre can be effectively illustrated.

The Sound Signature tool, which was developed as an extension to the Equator interface, shows how perceptually motivated visual mappings for timbre features can be used to guide and illustrate the process of sound design. The tool is based around a direct interaction paradigm, where the user records an audio signature and then uses the synthesis parameters to visually sculpt the timbre. The technique of offline audio rendering is used in the Sound Signature tool to provide instant visual representation of how synthesis parameters affect the timbre of the audio. In the context of Equator, the Sound Signature tool provides a perceptually meaningful representation of timbre that guides the use of the rest of the interface. The Sound Signature images also provide illustrative representations of Equator presets, and can be used to improve the process of preset selection.

The TimbreSphere tool was also developed as an extension to the Equator interface. TimbreSphere is aimed at performative exploration of the Equator parameter space. The tool provides a real-time data driven animation of the timbre as the user gesturally interacts with the underlying sound engine. By moving their hands in space, the user can explore the different timbral possibilities of the engine. Effectively, a complex virtual timbre space is provided for the user. Each of the dimensions of this space control multiple timbre features through complex mappings. The perceptual visualisation is used to represent how the various dimensions of timbre are affected by the user's gestures, in a perceptually meaningful way. The performative exploration of Equator's synthesis parameter space is guided by perceptually motivated representation of the underlying timbre space. This demonstrates how perceptually motivated visual mappings of timbre features can be used to influence interface design, and also to provide real-time representation of performative timbre variation.

Audio-visual performances have been described that were produced with the intention of using data-driven real time generative visualisations to accompany and draw attention to timbre variation as compositional structure. The music in these performances was centred around contrasting timbres and performative variation of timbres, making use of similar concepts found in electroacoustic music and spectromorphology. Stochastic motion simulators were used to provide visual representations of these timbral qualities and events. For each performance, specific mapping strategies were developed from acoustic timbre features to specific parameters of the stochastic motion simulation algorithms that were used. The performance at the Roundhouse in London was evaluated using questionnaires for both audience members and performers. The results of the audience questionnaire show that the respondents perceived connections between the visualisations and the audio, and also that the visualisations were descriptive of the audio. This suggests that the visualisations were effective in representing variations in timbre in real time. There was more uncertainty in the respondents as to the level of control they felt over the visualisations when interacting with the system. The performers questionnaire also shows that respondents felt a connection between the sounds and the visuals. There was variation in the level of control that they felt over the system, likely due to the differences in mapping strategies that were implemented for each performer. There was agreement that the visual responses were an important part of the performances and that they influenced the performances. This suggests that the system was used as an interactive performance tool as well as a passive real-time representation. The results of the questionnaires can only be used to evaluate the Roundhouse performance specifically, and cannot be generalised. Future studies would be required with larger numbers of participants and more

controlled environments in order to draw general conclusions. However, the visualisation system used at the Roundhouse can nonetheless be seen as a specific example of semantically-oriented timbre representation in the context of live performance.

The TimbreFluid system was developed as an experimental tool to try to characterise some general mappings between timbre features and properties of motion in stochastic motion simulators that can be applied across general contexts for use with different tools and algorithms. The TimbreFluid system is a specific tool that demonstrates the use of perceptually motivated mappings for the visual representation of timbre and temporal variation of timbre. However, in documenting and explaining the development process and motivations behind the mappings used, a generalised framework for the use of stochastic motion simulators for timbre representation has been developed. Some general visual properties have been categorised that are common to different stochastic motion simulation tools. General mappings have been proposed that link individual acoustic timbre features to these various visual parameters.

Chapter 5

Perceptual Timbre Representation

So far a number of individual experiments and systems have been described and discussed that investigate different visual mappings for different timbre features in different contexts and for different use-cases. This section brings together the results of experimentation and research-led development of novel systems and outlines a unified approach to perceptual timbre representation and visualisation. Building on the practice-based research described in the previous chapter, this chapter develops a theoretical formalisation of perceptual timbre representation. It provides guidelines for the development of such representation techniques that can be used by other practitioners.

This section categorises three types of perceptual representation and discusses the common corresponding use cases for each type of representation. They cover a combination of physically-oriented and perceptual semantically-oriented mappings. A discussion is given on the similarities and differences in constraints and requirements between developing different kinds of timbre representation tools. This chapter directly addresses the key aim of developing and formalising a perceptually motivated approach to timbre representation and visualisation.

The first section groups perceptual timbre representation into three categories. In the second section, these categories of representation are discussed in terms of the common use cases to which they are suited. The third section discusses the different requirements and constraints of each representation category. The fourth section discusses the issue of mapping and mapping strategies with reference to existing literature and to experiments and systems described earlier in the thesis. Various concepts involved in mapping timbre features to visual features are discussed including the level of abstraction, the concept of time, various visual parameter categories, and interactions between timbre, pitch and volume. Differences between representation categories are discussed in the context of these different mapping concepts.

5.1 Types of Representation

For the purposes of this discussion, timbre representation is categorised into three approaches: *identification*, *temporal* representation and *real-time* representation.

5.1.1 Identification

Timbre *identification* refers to the visual representation or identification of a single timbre category. In standard Western music notation this is equivalent to instrumentation. In the context of an audio production tool the timbre category is usually identified using the name of a track or virtual instrument (e.g. 'Piano,' 'Pad Synth', 'Drums'). Effectively, the task of timbre identification is to denote a particular timbre space (e.g the timbre space of a piano, or that of a voice).

The concept of timbre identification is linked to the traditional definition of timbre as the collective term for all of the aspects of an audio signal not related to the pitch or dynamics. As mentioned previously in section 2.1.4, this definition of timbre is inspired by the use of timbre perception in audio source identification. It is this kind of audio source identification that the *timbre identification* category of representation refers to.

In spectromorphological and musicological terms, the concept of *timbre identification* is related to the idea of seed structures that are combined and altered in various ways to produce composite timbral structures and events. The simple 'visual sound shapes' described by M. Blackburn (2011) identify simple sound objects with specific timbre features. The identification of simple sound types with individual characteristics can then form the basis for more complex forms of representation that indicate temporal variation, for example.

5.1.2 Temporal Representation

If timbre *identification* is about denoting a particular timbre space, or a specific point *within* a timbre space, then temporal timbre representation is the task of representing a particular temporal progression or movement through that space, from beginning to end. In Western musical notation an analogy is the use of articulation indicators (e.g. pizzicato, arco). These indicators can be seen as indicating particular regions of the timbre spaces afforded by different musical instruments. In musical production tools, progressions through timbre spaces are often represented using automation graphs which determine how the different features of a given virtual instrument change over time.

In his discussion of spectromorphology, Smalley (1997) categorises spectral events into gestures and textures and describes how they can be combined to produce motions and behaviours. These gestures, textures, motions and behaviours describe timbre in perceptual terms. They refer to perceptual impressions resulting in the listener from specific spectral events and temporal variations. As such, these spectromorphological concepts are directly linked to the temporal representation of timbre. M. Blackburn (2011) demonstrates how the visual vocabulary of spectromorphology and the concepts of spectral textures and behaviours can be transformed into illustrative graphical icons and representations. These graphical illustrations are temporal representations of timbre

that represent perceptual responses in the listener rather than direct aural features.

5.1.3 Real-Time Representation

Finally, *real-time* representation is the task of representing the moment-to-moment progression through a timbre space in real time, in direct response to audio. If the *temporal* representation shows an entire progression through the timbre space from beginning to end, then the *real-time* representation effectively goes 'inside' the progression and shows instantaneous moments as they occur, in a continuous animation. Similarly, real-time spectromorphological representation would be based around real-time generative animation of spectromorphological sound shapes such as those described by M. Blackburn (2011). The real-time representation is rarely used by commercial tools and applications. One analogous example is that of a real-time FFT analyser such as the 3D spectrograph in the DMG Dualism plugin, or real-time audio visualisers in audio media player applications such as the iTunes visualiser. However, most systems for *real-time* timbre representation exist in the context of audio-visual performance and art installations and have been developed for very specific contexts, as discussed previously in section 2.5.4.

5.2 Use Cases

Each of the three categories of timbre representation are suited to specific scenarios, user goals, and applications.

5.2.1 Identification - Selection, Grouping, Arrangement & Production

Identification of timbre categories is important in the selection and grouping of sound types, and in the arrangement of sound types in both a compositional and production context. For example, in a tool for sound design the user goal of preset selection makes use of timbre identification. Presets are often grouped by type (e.g. strings, synths, drums) and often have names that indicate ceertain timbral qualities (e.g. 'smooth pad'). In 'traditional' composition and arrangement, specific sections are assigned to specific timbre groups. In an audio production tool such as a digital audio workstation, specific tracks need to be identified by their timbral features. As mentioned previously in section 5.1.1, this is usually indicated through the name and/or using some form of abstract colour coding strategy.

As discussed in the previous section, M. Blackburn (2011) has shown how timbre identification plays an important role in spectromorphological sound sculpting. Timbre identification forms the basis of visual sound shape selection, and these individual sound shapes – or objects – are then combined and altered in different ways.

5.2.2 Temporal Representation - Arrangement, Editing & Scoring

The temporal representation of timbre is important for both graphical scoring of a performance and for timeline editing of audio data and parameters. In a musical score or any kind of temporal representation of musical and performance data, indication must be given of how the timbre should – or does – evolve and change over time. In the context of electroacoustic works, multiple forms of timbre notation techniques have been developed. They are used for performance, analysis and appreciation of electroacoustic music. In musical production tools, the overall progression through a given timbre space needs to be searchable such that particular regions can be isolated for editing and tweaking.

Temporal timbre representation can also be used as a deeper form of timbre identification. It can be used to categorise and contrast timbral features of various instruments in a multi-track production environment. The representation of temporal variation of timbre can also be used to distinguish between individual sounds with similar characteristics that wouldn't necessarily be immediately separable using a simpler timbre identification method (e.g. instrument labeling or abstract colour-coding).

5.2.3 Real-Time Representation - Performance

Real-time representation of timbre is used in performance settings to visualise the moment-to-moment variation of timbre features and can be used to draw attention to the use of timbre variation as a compositional device. This has been explored by the various audio-visual performances discussed in the previous chapter.

The real-time representation of timbre can also be used in the design of new interfaces for performative timbre manipulation. It can be used for performative feedback in order to guide performance and exploration, as demonstrated by the TimbreSphere system. By representing timbre as a sound object, the mappings can be reversed such that changes to the object produce changes in the audio. This provides a sense of direct interaction, where the object is being sculpted or manipulated in some way in order to produce timbral effects in a performance context.

5.3 Requirements & Constraints

5.3.1 Accuracy

The accuracy of the timbre representation will be determined by the number and variation of audio features that are used. Peeters et al. (2011) provide a set of minimum requirements that should be met in order to obtain a minimally sufficient measure of the timbre of a signal. However, the level of accuracy will also be constrained by the type of representation. In real-time representation, for example, the global timbre features will be poorly represented, since the visualisation algorithm only has access to instantaneous audio information. Timbre identification techniques will be, by definition, less detailed and accurate than either temporal or real-time representations. Since

temporal techniques combine both instantaneous and global features, they have the potential to provide the most accurate representations.

5.3.2 Identification

The most common way to represent the entire timbre space afforded by a certain sound or category of sounds is to refer to instrumentation ('piano,' 'synth'). In some contexts such as Western classical music this may be sufficient, particularly when the instrumentation consists of well-known instruments with instantly recognisable timbres. However in other contexts such as electronic music and production, or electroacoustic music, and other forms of contemporary music, the use of simple labels to denote timbre categories can be problematic. As a simple example in the case of music production, there can often be multiple instances of different types of instrument (e.g. 'piano 1, piano 2' or 'bass synth 1, bass synth 2'). The use of labels to categorise more subtle differences can be inefficient and often leads to long, specific naming conventions based on semantic language (e.g. 'soft gliding pad synth', 'hard crunchy synth bass'). This issue often comes from the fact that many new digital instruments afford very wide-ranging timbre spaces, especially in the case of user-specified architectures. Naming a preset then becomes a process of describing the kind of sound that the current configuration of the sound engine produces. In this thesis a technique referred to as 'synthesis visualisation' has been explored which addresses this issue. Synthesis visualisation invloves the development of visual representations of the sound engine configuration using a specific visual synthesis algorithm.

In terms of graphical representation, the key requirement of perceptual timbre identification is to accurately depict the kind of timbre qualities that are afforded by the particular timbre space of the sound or sound category that is being represented. Similarly, perceptual timbre identification might be used to represent a static point within a timbre space (as in the case of spectromorphological sound shapes). A major constraint is that the visual representation needs to be very compact and instantly recognisable and distinguishable from other category representations. The influential 'kiki bouba' experiment (Kohler 1929) has shown that spiky shapes are very effective representations of specific types of sound ('kiki') in comparison to round shapes, which are more effective in representing different sound qualities ('bouba'). This is effectively the key task for perceptual timbre identification: to make use of some kind of language – whether visual, abstract, text-based or other – in order to effectively categorise and distinguish different sound types (timbre spaces) with reference to semantic descriptors.

5.3.3 Temporal Representation

One of the main constraints for temporal representation methods is the need to represent time and temporal progression. The way in which time is represented defines the overall representation. Time is conventionally represented linearly from left to right in many other forms of media including musical scores, audio production tools and timeline editors. It therefore makes sense that examples of temporal timbre representation methods and timbre notation methods also represent time linearly, from left to right. Given this initial constraint, the temporal representation method must make use of some form of visual mapping strategy – either abstract and symbolic or low-level and descriptive – whereby visual features along the timeline indicate how the timbral

characteristics evolve over time. Perceptual temporal timbre representation makes use of semantic descriptors and visual analogies in order to depict the evolution of timbre characteristics.

Such representation of a progression through a timbre space should highlight specific regions of the timbre space – and movement between them – over time. It should show the contrasts and similarities between different regions of the timbre space. As well as charting the temporal variation of timbral qualities, the representation should also be readable such that the timbre of every instant is represented. For example, the use of parameter automation graphs in audio production tools shows how different parameters vary but doesn't show how the timbre sounds overall, from moment to moment. Rather, the serial inspection of multiple parameter values is required, or the aural preview of particular regions. One of the main requirements of the temporal timbre representation is to provide a readable representation of the trajectory of a sound through a given timbre space such that relevant regions are immediately identifiable for inspection and such that contrasting regions can be easily visually compared.

5.3.4 Real-Time Representation

Like temporal representation, the main constraint for real-time timbre representation also concerns time. However, the constraint in the case of real-time representations is that only small windows of time are accessible at any given moment. Real-time representations are procedural in that new information is constantly being analysed in order to drive the representation to the next state. As such, real-time timbre representation methods have similar constraints to other real-time applications such as games and musical instruments. They need to be computationally efficient and inexpensive in order to reduce jitter and latency.

The key requirements of real-time timbre representation are reactiveness and synchronisation. Visual events need to be completely synchronised with aural events to maintain the perceptual link between aural and visual stimuli. In real-time and performance contexts the concept of *motion* is one of the most important issues, and Smalley's categories of motion are particularly relevant. As discussed prevously, real-time representations are equivalent to 'going inside' the temporal representation and experiencing the timbral progression from moment to moment. In the animation, the motion *between* these moments needs to be smooth and should be directly guided by the aural features.

5.4 Mapping

All three categories of representation are dependent on a mapping strategy that defines how particular timbre features are represented.

5.4.1 Abstraction

Representation strategies for timbre can range from direct physical mappings (e.g. spectral centroid to brightness) through to completely abstract categorisations (e.g. the use of colours in

DAWs to categorise different types of audio track). Identification of timbre is the representation category that can make use of the most abstract mapping strategies. The key constraint is that consistency is maintained such that timbres and timbre groups can be easily distinguished. Temporal timbre representations can also make use of abstract mappings but a key constraint is that they provide some representation of trajectory or progression over time. For example in graphical scores, gradients are often used to indicate directional spectral movement of energy. In real-time representation, accurate perceptually meaningful representation of moment-to-moment evolution of timbral features is best achieved through the use of more specific mappings.

5.4.2 Time

While temporal and real-time representations are dependent on temporal variation and evolution, timbre identification has no such constraint. Timbre identification is the task of depicting an overall impression of the possible timbre space. Mapping strategies can therefore be designed without reference to time. The key difference between temporal representation and real-time representation in the context of time is that temporal representations are created 'offline' with direct access to all of the audio or performance data at the outset. For example, in the Sound Signature technique the entire visualisation is rendered and then a second visual rendering process is implemented where a global blurring is applied to the signature depending on the spectral centroid value at each individual time point. Temporal representations can therefore include macro-level events that occur over long time periods. With real-time representation, the only information available is the instantaneous audio feature data, and possibly some buffer of past data values (for example in the case of smoothing of values). Real-time representation therefore depends on specific physically or perceptually meaningful mappings such that each time slice of audio data produces a coherent visual frame. The *motion* between points in the timbre space thus emerges directly as a quick steady sequence of adjacent points in the timbre space, as in any other procedural animation process.

5.4.3 Colour & Luminance

A perceptual association for which there is a very substantial body of evidence from existing research is the association between spectral centroid and perceived brightness of a sound (Beauchamp 1982; De Poli and Prandoni 1997; Zacharakis, Pastiadis, and Reiss 2014). A preference has been shown in listeners for the mapping from spectral centroid to physical brightness in a visual representation (Berthaut, Desainte-Catherine, and Hachet 2010). The Comparisonics waveform display technique maps low frequency content to blue colour and high frequency content to red colour. Research has also suggested a tendancy in listeners to pair audio-visual stimuli according to this mapping (Adeli, Rouat, and Molotchnikoff 2014). The results of the Sound Signature user survey discussed earlier support the claim that an inverse mapping from spectral centroid to colour hue is generally preferred in listeners. This thesis and the systems described within have made the case for an inverse mapping from spectral spread to colour saturation. This mapping is physically motivated. If spectral centroid is used to sample a particular point in the colour spectrum (using the hue value) then the spectral spread can indicate how concentrated that colour should be. In essence the colour spectrum is being sampled according to salient features of the audio spectrum.

5.4.4 Texture

Another audio-visual association that has been suggested from multiple studies is the association between aural noise and visual noise (texture) (Berthaut, Desainte-Catherine, and Hachet 2010; Giannakis 2006). The spectral flatness can be used as an indication of the noisiness of a signal (high flatness) as opposed to tonality (low flatness). This can then be mapped to the level of granularity in a visual texture for an effective representation of the noise content of a signal. The harmonicity of a signal can also be effectively visualised using texture. Sound Mosaics is one example in which the 'sensory dissonance' is visualised using texture regularity (Giannakis 2001). Sensory dissonance relates to the deviation of frequency partials from pure harmonicity (i.e. integer multiples of the fundamental frequency).

These texture mappings have been demonstrated by the systems described in the previous chapter. In Sound Signature visualisations, the flatness controls the level of visual noise, implemented using a colour distortion technique. In the TimbreFluid example system, the spectral flatness controls the vorticity scale, which means that noisier signals produce more fine-grain textures in the fluid. Similarly, in the TimbreSphere system, the spectral flatness is mapped to the frequency of deformation around one polar axis of the sphere. TimbreSphere also makes use of texture to represent harmonic characteristics. The harmonic energy ratio controls the bump mapping amount, and inharmonicity controls the bump mapping granularity.

5.4.5 Shape

Multiple studies have shown correlations between visual shapes and distinct timbre identities. The 'kiki bouba' experiment is an influential example (Kohler 1929). Also more recently, Adeli, Rouat, and Molotchnikoff (2014) showed a tendency in participants to associate distinct timbres with certain shape types. As discussed previously in section 2.4.2, these studies use distinct timbre categories and do not quantify differences between timbres, which causes difficulties in making use of the results in the development of mapping strategies. An exception is the case where only entire timbre categories need to be represented. For this reason, visual shape representations are suited to the task of timbre identification. Parise and Spence (2012) showed that participants were able to accurately pair audio-visual stimuli where the low-level waveshape was used to determine the visual shape representation. Participants were played aural stimuli consisting of sinusoidal tones or square-wave tones. The visual stimuli were rounded curves and more angular linear forms. Participants showed a tendency to associate the sinusoidal tones with the rounded curves and square-wave tones with angular forms. This suggests a low-level quantifiable mapping from wave shape to visual shape representation that would be understandable not just by those with knowledge of sound synthesis. Such a mapping could be used to determine the initial seed shape – or structure – of a visualisation, which could then have visual effects applied, much in the same way digital wavetable synthesis consists of original waveform shapes that then have audio effects applied.

The use of low-level waveform shape to generate visual shape representations has been explored by Putnam (2014) who used rotation of sinusoidal waves in 3D space to generate procedural shapes. This use of low-level shape features of the audio to determine visual shape and form in the visual representation could be a good starting point for many mapping strategies, such that texture, colour and motion then build upon the existing structure. The technique of synthesis visualisation

described earlier in the thesis makes use of visual shape to semantically represent the sound qualities of a preset (sound engine configuration). These visual sound shapes can then be used for preset identification and search and retrieval. TimbreSphere also makes use of varied 3D shapes through spherical deformation, which are defined by the spectral centroid and spectral flatness.

5.4.6 Motion

Motion is an important issue in temporal and real-time representations of timbre, which in effect represent motions through a timbre space, from different perspectives (as discussed previously). Smalley (1997) demonstrates in his discussion of spectromorphology the idea that motion in the spectrum can be used to characterise different timbral qualities. Quantitative psychological research carried out by Elliott, Hamilton, and Theunissen (2013) supports this idea. Their results showed that all semantic descriptors from a VAME study were correlated with temporal fluctuation in the spectrum at different frequency bands. Unsurprisingly, the differences in mappings for motion between temporal and real-time representations are similar to the differences concerning time. In temporal representations, 'motions' and gestures are directly accessible from beginning to end. Macro-level mappings can therefore be constructed in order to categorise and distinguish individual motions. For example the entire signal could be analysed in order to extract transients and silent periods, such that individual sound snippets could be further analysed and distinguished. In real-time representations, the idea of motion must emerge from instantaneous mappings that work on a frame-to-frame basis, as described earlier. As an example, in the TimbreSphere demo system, the root-mean-square amplitude is mapped to the magnitude of spherical deformation. The object therefore animates in response to fluctuations in amplitude.

5.4.7 Mass / Density

Multiple studies into verbal descriptors of timbre have identified mass or fullness as a salient independent semantic descriptor of timbre. Fullness has been shown to be correlated with fluctuation in low-end energy in the spectrum (Alluri and Toiviainen 2010). In a separate study, the 'thin' descriptor was found to correlate with lack of energy in the low-end of the spectrum (Disley, Howard, and Hunt 2006). Similarly, Helmholtz and Ellis (1875) described fullness as the prime tone outweighing the upper partials. TimbreFluid makes use of a mapping from low frequency energy ratio to viscosity level in the fluid simulation, meaning that sounds with fuller low-end energy produce more dense fluid with slower, more resistive movement.

5.4.8 Interaction With Pitch & Volume

As discussed earlier there are multiple studies showing that perception of timbre and pitch and perception of timbre and volume can interact in certain contexts. For pitch, studies show that the pitch context affects the extent to which timbre is attended to (Krumhansl and Iverson 1992), and also that the level of attention given to timbre and pitch both depend on the extent to which the other is in flux (P. G. Singh and Hirsh 1992; Caruso and Balaban 2014). In visual representation systems, these findings suggest that pitch should be mapped to a particular visual feature that varies independently of other features that represent timbre. For example, pitch might be mapped

to the height of a visual object whereas timbre is mapped to the surface characteristics of that object. Interaction between the perception of timbre and and the perception of loudness has also been demonstrated in listeners (Melara and Marks 1990). Amplitude should therefore also be mapped to some other independent feature. Often this independent feature is the size. In terms of visual mappings the amplitude can act as a weighting on the mappings. For example, if spectral centroid is mapped to brightness, this can be weighted by the amplitude such that brightness is controlled by the spectral centroid value multiplied by the amplitude value.

5.5 Conclusion

This chapter has brought together results from experimental psychology and audio-visual user studies as well as concepts from acoustics and musicology (spectromorphology) in order to present general approaches to the perceptual representation of musical timbre. The approaches have grounding in both perceptual and physical qualities of timbre. Three categories of timbre representation are discussed. Temporal and real-time timbre representation are used to represent the variation of timbre over time, and the (perceptual) movement or 'motion' of a sound through a particular timbre space. Temporal timbre representation shows this motion in its entirety whereas real-time representations are procedural animations driven directly by the timbre features that give a moment-to-moment perspective of the motion. Timbre identification is the task of representing the overall timbre space inside which a particular motion may occur. In less theoretical terms, timbre identification is the task of showing particular sound categories.

Each of the three representation categories are suited to different use cases and have different requirements and constraints. Timbre representation systems are defined by mapping strategies – collections of mappings from individual timbre features to visual features. The development of mapping strategies has been discussed in detail in this chapter. Both the level of abstraction involved in the mappings and the notion of time and causation affect the different categories of representation in different ways. A collection of visual feature categories has been discussed and some general mappings for the visual representation of timbre features have been proposed with reference to the specific representation category they relate to.

Chapter 6

Conclusion

6.1 Main Aims

The key aim of this thesis has been to develop a perceptually motivated approach to timbre representation. This key aim has been motivated by concepts in electroacoustic music and spectromorphology and the use of visual and behavioural vocabulary to describe and formalise the musical variation of timbre over time. Similarly, the visual vocabulary of timbre description identified by various studies in experimental psychology, combined with potential acoustic correlates for specific visual qualities, has been a motivating factor. This has influenced the development of data-driven perceptually motivated timbre representation techniques that can be applied in different contexts.

This perceptually motivated approach has been applied in two specific contexts. Firstly, interface design in digital tools for timbre creation and manipulation, and secondly in the context of real-time performance.

6.2 Summary

Chapter 2 has highlighted some of the key issues with the standard definition of timbre and provided an alternative definition. The perception, description and measurement of timbre have been discussed. The concept of a timbre space has been introduced and the key perceptual, semantic and acoustic descriptors of timbre have been identified. Commonalities between these descriptors have been used to put forward the case for data-driven perceptual timbre representation techniques. Existing examples of similar representation methods have been reviewed.

Chapter 3 has described various experiments and experimental tools that were developed in order to investigate the mapping of acoustic timbre features to semantic timbre descriptors. A preliminary study using 3D animated visual stimuli identified variability in user preferences when multiple features were in flux. An experimental mapping tool has been described that investigates

the use of synthesis parameters to drive visualisation. It can be used to produce abstract illustrative representations similar to spectromorphological sound shapes. A novel data-driven temporal timbre representation technique has been introduced and evaluated using a web-based interactive user survey. Results have highlighted generalisable preferences for specific mappings.

Building on the results and techniques from Chapter 3, Chapter 4 has described various systems and performances that were developed as working demonstrations of the use of perceptual timbre representation in various contexts. They are all based on the use of a feature extraction tool that has been developed for use in real-time contexts. The Sound Signature demonstrates the use of perceptual timbre representation in the context of sound design and audio production. Three audio-visual performance events have been described which make use of perceptual timbre representation in the context of live musical performance. Evaluation of one such performance has shown a general agreement in audience members and performers as to the presence of a connection between audio and visuals. Results also suggest minor agreement that the visuals were descriptive of the audio. A real-time timbre representation tool has been described that was developed for use in such contexts.

The Sound Signature tool and the Roundhouse performance, both described in Chapter 4, have provided demonstrations as to how perceptual timbre representation can be implemented and used in different contexts. Chapter 5 has brought together the results of this practice-led development and formalised a general framework for perceptual timbre representation.

6.3 Key Results and Discussion

Through the development of novel timbre representation techniques inspired by results from experimental studies, a collection of general approaches to perceptual timbre representation have been developed. These were presented and discussed in the previous chapter. Three key categories of representation are covered: identification, temporal representation and real-time representation. These categories can be understood with reference to the concept of timbre space. Timbre *identification* is the task of representing an entire timbre space. *Temporal* timbre representation is the task of representing a specific temporal progression through a timbre space. *Real-time* timbre representation is the task of representing the instantaneous timbre characteristics of a sound from moment to moment during a progression through a timbre space.

This thesis provides examples of novel representation techniques in each category. The technique of 'Synthesis Visualisation' was explored mainly for the task of timbre identification. It involves the use of congruent visual and audio synthesis and a mapping layer between them. This facilitates the perceptual representation of the synthesis engine's configuration. The Sound Signature technique was developed as a temporal timbre representation method. It makes use of an offline audio rendering technique such that new audio data can be analysed and represented instantaneously. Multiple real-time timbre representation techniques have been explored in this thesis. Each of them involves real-time tracking of acoustic timbre features and subsequent mapping to visual properties.

In the previous chapter, which presents a collection of general techniques for perceptual timbre representation, the main issues involved in the development of mapping strategies were discussed and related to the three different categories of representation. The Sound Signature user survey

discussed earlier in this thesis has provided important results that can be used in the development of mapping strategies for perceptual timbre representation. It provides further confirmation for the growing body of evidence that the mapping from spectral centroid to visual brightness is favoured in listeners. The study also showed that the use of colour and texture to represent timbre was favoured over the absence of colour and texture mappings for timbre. In particular, the study highlighted the general preference in listeners for an inverted mapping from spectral centroid to colour hue (blue-to-red for low-to-high frequency content) over a direct mapping (red-to-blue).

In the context of live performance, real-time timbre representation methods are most appropriate. Multiple systems and methods for real-time timbre representation have been explored in this thesis in order to address the ways in which perceptual correlations can be used in real-time live performance contexts. The audience and performer questionnaires evaluating the Roundhouse performance show that real-time semantic timbre representation was successfully used in order to provide generative visual accompaniment to musical performances, and to draw attention to timbre variation as a compositional and performative device. In particular, the performances have demonstrated the use of stochastic motion simulation techniques for real-time timbre representation. General approaches to mappings from acoustic timbre features to system parameters have been presented. The use of such mappings in the various performances described in the thesis shows how semantic and acoustic correlations of timbre can be drawn upon in order to produce general real-time timbre representation systems for use in live performance contexts.

In the context of interface design, the development of the EMapper tool and the conception of the 'Synthesis Visualisation' technique has demonstrated that correlations between semantic and acoustic qualities of timbre can be used in the development of perceptual timbre identification methods in digital tools. The Sound Signature tool could potentially be used in applications as a way to bridge the gap between core-language and task-language. This would require further studies. The novel 'offline audio rendering' method developed as part of the Sound Signature tool affords real-time visual rendering, where manipulation of synthesis parameters produces immediate semantic feedback in the Sound Signature image. In this way, the Sound Signature technique can be used to provide visual representation of the resulting effects that the various synthesis parameters of the digital tool have on the resulting audio.

6.4 Contributions & Implications

This thesis has provided multiple demonstrations of how perceptual timbre representation can be used effectively in various contexts. It has brought together the results of experimental, research-led development in order to formalise the concept of perceptual timbre representation and provide a general framework for the development of such representation techniques.

Perceptual timbre representation is based on the use of semantic correlates of acoustic timbre features in order to provide visual representation of the high-dimensional timbral characteristics of an audio signal. As such, it facilitates the comparison of different timbres with respect to particular physical attributes and qualities (audio features). This contrasts with the standard definition of timbre as a collective measurement of similarity between sound characteristics that do not relate to pitch or loudness. This suggests that the current definition of timbre as:

> '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'

could be improved by making direct reference to the spectral and dynamic content of audio and its variation over time. For example, timbre may be better understood as:

'The spectral distribution of the audio and its evolution over time, as well as the dynamic variation.'

This thesis has demonstrated the use of stochastic motion simulators for real-time timbre representation in numerous live performance contexts and has shown how perceptual timbre representation can be used as an important accompaniment or central feature of a live audio-visual performance. Other live performances, particularly in the area of electroacoustic music, could make use of such techniques in order to develop a unified visual language of performative timbre representation. New digital interfaces for performative timbre manipulation could also make use of the real-time perceptual timbre representation methods and mapping techniques presented in this thesis. Real-time timbre representation could also potentially be used to develop direct interaction systems where the visual representation is directly altered in order to manipulate the timbre. This kind of interaction is simulated by the TimbreSphere system through the use of visual representation that is behaviourally related to the gestural input.

The Sound Signature technique is a novel perceptual timbre representation method that can be used in digital interfaces for timbre creation and manipulation. The Sound Signature survey has highlighted general preferences for specific visual mappings of timbre features in the context of temporal timbre representations. In particular the preference for an inverse mapping from spectral centroid to colour hue is an important novel contribution. This result builds on previous studies that have shown associations between specific opposing colours (e.g. blues and greens vs. reds and yellows) and specific opposing timbre labels (e.g. harsh vs. soft). The identification of an association between a specific audio feature (spectral centroid) and a specific colour parameter (colour hue) provides an important contribution to defining a universal way in which timbre can be effectively represented in digital tools. The integration of the Sound Signature tool with the existing Equator interface involved the development of the 'offline audio rendering' technique, which is used to optimise responsiveness in the Sound Signature tool. This is a technique which can be used in the development of other interfaces. In particular it can be used for immediate rendering of audio data for display in response to manipulation of synthesis parameters. The Sound Signature technique, combined with the various mappings discussed throughout the thesis, can be used to implement perceptual timbre representation methods, which could potentially bridge the gap between core language and task language in digital interfaces.

6.5 Limitations

The methodology of this thesis has mainly been focused on the experimental development of novel tools and techniques. These tools and techniques have made use of existing semantic and acoustic correlations of timbre. They have been developed and demonstrated as working examples of how such correlations can be put to use in novel techniques and systems in specific contexts. Although novel user evaluation has been gathered in the context of Sound Signature mappings

and representations, there are numerous areas in which further evaluation and user testing would be beneficial. The Sound Signature study has highlighted a preference for an inverted mapping from spectral centroid to colour hue, since two of the mapping strategies feature the centroid-hue mapping with opposing polarities, and one is preferred over the other. However, it is difficult to obtain similar evidence for the preference of other mappings (such as spectral flatness to visual noise, or spectral spread to colour hue) since these mappings are grouped together in the mapping strategies, and the preference of one strategy does not necessarily indicate the preference of each individual mapping in that strategy.

User evaluation could be carried out of the visualisations used in the TimbreSphere system in order to qualify or extend existing results. Similarly, user evaluation could be carried out on the TimbreFluid system by examining user preferences for the use of particular fluid simulation parameters in order to represent certain timbre features. Again, the results of such a study could then be compared with existing results from previous studies.

The preliminary user study described in section 3.1 featured a small number of participants, of whom few had experience with audio editing or sound design software. Future studies would benefit from larger participant pools involving participants with experience in sound design and music technology such that comparisons can be drawn. One of the key aims of the Sound Signature study was to fulfill these requirements. However, there are ways in which the Sound Signature survey could be improved and developed. A similar study could be implemented that uses other such temporal timbre representation techniques such as the Comparisonics coloured waveform display, and draws comparisons between these in terms of user preferences. The Sound Signature survey was based around user preferences, but a similar survey could be implemented which is designed to test the effectiveness of certain mappings over others, by involving the selection of a *correct* representation from a number of possible representations, in a similar way that Grill and Flexer (2012) implemented their online survey. The *correct* representation in this context would refer to the representation that has been generated using the corresponding audio stimulus for that question.

As is the case for the various systems described in this thesis, the visualisation techniques and visual elements used in the various performances were influenced by existing results and suggested perceptual correlations. As such they demonstrate the application of these correlations in the context of live performance. Again, there is potential to implement more formal evaluation in such live performance contexts. For example audience evaluation could be gathered on performances with and without visual elements in order to somehow quantify the ways in which visual representations alter audience perceptions in such contexts.

The experimental development and experimentation in this thesis has made use of spectral descriptors and harmonic descriptors that have been used in previous work. The level of accuracy of a representation is constrained by its context, as explained in section 5.3.1. Temporal techniques such as the Sound Signature technique, by representing both global and instantaneous features, are the most effective way to provide an accurate representation. The audio features used in the Sound Signature technique give a good account of the timbre of a signal, according to the minimal requirements suggested by Peeters et al. (2011). However, there are many other features that could be considered, some of which measure similar qualities as those used in this thesis, and some which measure different aspects. In particular the modulation power spectrum has been shown to provide numerous salient indicators of timbre (Elliott, Hamilton, and Theunissen 2013). MFCCs have also proven useful in the modeling of both instrumental and vocal timbres (Eronen 2003; Heittola, Klapuri, and Virtanen 2009; Chudy and Dixon 2013; Tsai and Wang 2006). Since the MFCCs of

a signal constitute a multi-dimensional feature, an interesting area of research would be to make use of the Wekinator tool (Fiebrink, Trueman, and P. R. Cook 2009) to associate them with a visualisation algorithm, in a similar way that the Wekinator tool was used in the Roundhouse performance (see section 4.5). The discrete wavelet transform could also be an interesting feature to explore. It provides high time resolution and low frequency resolution for high frequencies and high frequency resolution and low time resolution for low frequencies (Tzanetakis, Essl, and P. Cook 2001) and so could be useful in obtaining a good measure of the frequency distribution of the spectrum.

6.6 Future Research

Many of the tools and techniques discussed in this thesis could be used as the basis for specific user tests and experimentation. The EMapper tool was developed in order to facilitate complex mappings from audio synthesis parameters to visual synthesis parameters. The resulting visualisations could form the basis of a user study that investigates users' abilities to identify particular presets and timbres visually. The TimbreSphere system could be used in order to examine user preference for specific gestures over others in the control of synthesis parameters.

The Sound Signature study measured user preferences for visual mappings. Future studies should examine the effectiveness of these mappings in the context of goal-oriented tasks, with standard amplitude waveforms as a comparison. The survey was implemented as an interactive web application which allowed a large diverse participant pool to be reached. The use of the WebGL and Web Audio APIs facilitate the development of complex audio visual applications. The technique of interactive web-based user tests and surveys could therefore be extended in order to examine user preferences for – and effectiveness of – more complex mappings. For example, the technique of audio-driven generative modeling and rendering used in the TimbreSphere system could be implemented in an interactive web app. This would allow a large user evaluation of complex 3D audio-visual mappings.

Most of the tools and techniques discussed in this thesis used the ROLI Equator synthesis engine and interface as a starting point. As a result, they are most suited to digitally synthesised audio. Perceptual timbre representation could be put to use in other digital tools and in combination with other synthesis techniques. Interfaces for concatenative synthesis, for example, could benefit from the use of perceptually-motivated visualisation of individual grains. This could facilitate the detailed representation of individual points within an interactive timbre space. Like the Comparisonics coloured waveform display technique, the Sound Signature technique has many applications across different sound design and audio editing contexts. It could be used effectively in DAW environments to optimise the processes of track arrangement and organisation as well as audio editing and region selection. The thesis has also explored perceptual timbre representation in the context of the human voice, and live performance. It would be interesting to explore other types of audio signal such as traditional musical instruments or sound effects.

Both the Sound Signature and TimbreSphere systems simulate a form of direct interaction, as manipulation of synthesis parameters produce immediate visual results. In the case of the TimbreSphere system this is enhanced by using visualisations that are behaviourally related to the gestural input. Direct interaction could be more directly implemented however. For example the Sound Signature tool could be developed further such that direct manipulations of the Sound

Signature representation itself are used in order to sculpt the audio. Similarly, the TimbreSphere system could be altered such that the input involves the actual deformation of a sphere in virtual reality or augmented reality.

Although this research has been carried out in the context of audio editing, sound design, music production and performance, there are potential applications of perceptual timbre representation in other areas. Particularly in game development, virtual reality, and other interactive media, generative real-time perceptual timbre representation could be used in the creation of immersive audio-visual worlds where audio input has direct perceptually salient representation in the visual environment.

This thesis has attempted to lay out the basis for a perceptually motivated approach to timbre representation in numerous different contexts. The pervasive use of such timbre representation in digital tools and in real-time performance contexts could lead to the development and widespread use of a common visual language for timbre description.

Bibliography

Adeli, M., J. Rouat, and S. Molotchnikoff (2014). "Audiovisual correspondence between musical timbre and visual shapes". In: *Frontiers in human neuroscience* 8.

Agostini, G., M. Longari, and E. Pollastri (2003). "Musical instrument timbres classification with spectral features". In: *EURASIP Journal on Applied Signal Processing* 2003, pp. 5–14.

Alexander-Adams, S. (2015). "A Flexible Platform for Tangible Graphic Scores". PhD thesis. University of Michigan.

Alluri, V. and P. Toiviainen (2010). "Exploring perceptual and acoustical correlates of polyphonic timbre". In: *Music Perception* 27, pp. 223–241.

Battier, M. (2015). "Describe, Transcribe, Notate: Prospects and problems facing electroacoustic music". In: *Organised Sound* 20.01, pp. 60–67.

Beal, A. L. (1985). "The skill of recognizing musical structures". In: *Memory & Cognition* 13.5, pp. 405–412.

Beauchamp, J. W. (1982). "Synthesis by spectral amplitude and 'Brightness' matching of analyzed musical instrument tones". In: *Journal of the Audio Engineering Society* 30.6, pp. 396–406.

Bellemare, M. and C. Traube (2005). "Verbal description of piano timbre: Exploring performer-dependent dimensions". In: *Digital Proceedings of the 2nd Conference on Interdisciplinary Musicology (CIM05)*. Observatoire interdisciplinaire de création et de recherche en musique (OICRM) Montreal, QC.

Berger, K. W. (1964). "Some factors in the recognition of timbre". In: *The Journal of the Acoustical Society of America* 36.10, pp. 1888–1891.

Bernays, M. and C. Traube (2011). "Verbal expression of piano timbre: multidimensional semantic space of adjectival descriptors". In: *Proceedings of the International Symposium on Performance Science (ISPS2011)*. European Association of Conservatoires (AEC) Utrecht, Netherlands, pp. 299–304.

Berthaut, F., M. Desainte-Catherine, and M. Hachet (2010). "Combining audiovisual mappings for 3d musical interaction". In: *International Computer Music Conference*, p–100.

Bismarck, G. von (1974). "Timbre of steady sounds: A factorial investigation of its verbal attributes". In: *Acta Acustica united with Acustica* 30.3, pp. 146–159.

Blackburn, A. and J. Penny (2015). "Timbral Notation from Spectrograms: Notating the Un-Notatable?" In: *Proc. of the Int. Conf. on New Tools for Music Notation and Representation TENOR*, p. 218225.

Blackburn, M. (2011). "The Visual Sound-Shapes of Spectromorphology: an illustrative guide to composition". In: *Organised Sound* 16.01, pp. 5–13.

Blinn, J. F. (1977). "Models of light reflection for computer synthesized pictures". In: *ACM SIGGRAPH Computer Graphics*. Vol. 11. 2. ACM, pp. 192–198.

Bregman, A. S. and S. Pinker (1978). "Auditory streaming and the building of timbre." In: *Canadian Journal of Psychology/Revue canadienne de psychologie* 32.1, p. 19.

Caclin, A. et al. (2005). "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones". In: *The Journal of the Acoustical Society of America* 118.1, pp. 471–482.

Carifio, J. and R. Perla (2008). "Resolving the 50-year debate around using and misusing Likert scales". In: *Medical education* 42.12, pp. 1150–1152.

Carroll, J. D. and J.-J. Chang (1970). "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition". In: *Psychometrika* 35.3, pp. 283–319.

Caruso, V. C. and E. Balaban (2014). "Pitch and timbre interfere when both are parametrically varied". In: *PloS one* 9.1, e87065.

Chi, T., Y. Gao, et al. (1999). "Spectro-temporal modulation transfer functions and speech intelligibility". In: *The Journal of the Acoustical Society of America* 106.5, pp. 2719–2732.

Chi, T., P. Ru, and S. A. Shamma (2005). "Multiresolution spectrotemporal analysis of complex sounds". In: *The Journal of the Acoustical Society of America* 118.2, pp. 887–906.

Chudy, M. and S. Dixon (2013). "Recognising Cello Performers Using Timbre Models". In: *Algorithms from and for Nature and Life*. Springer, pp. 511–518.

Corbett, R. et al. (2007). "Timbrefields: 3d interactive sound models for real-time audio". In: *Presence* 16.6, pp. 643–654.

Creasey, D. P. (1998). "An exploration of sound timbre using perceptual and time-varying frequency spectrum techniques." PhD thesis. University of York.

Dannenberg, R. B. (1993). "Music Representation Issues, Techniques, and Systems". In: *Computer Music Journal* 17, pp. 20–30.

De Boer, E. (1976). "On the "residue" and auditory pitch perception". In: *Auditory System*. Springer, pp. 479–583.

De Cheveigné, A. and H. Kawahara (2002). "YIN, a fundamental frequency estimator for speech and music". In: *The Journal of the Acoustical Society of America* 111.4, pp. 1917–1930.

De Poli, G. and P. Prandoni (1997). "Sonological models for timbre characterization*". In: *Journal of New Music Research* 26.2, pp. 170–197.

Deutsch, D. (1984). "Psychology and Music". In: *Psychology and its Allied Disciplines*. Ed. by M. H. Bornstein. Hillsdale: Erlbaum, pp. 155–194.

Di Scipio, A. (1994). "Micro-time sonic design and timbre formation". In: *Contemporary Music Review* 10.2, pp. 135–148.

Disley, A. C. and D. M. Howard (2004). "Spectral correlates of timbral semantics relating to the pipe organ". In: *Speech, Music and Hearing* 46, pp. 25–39.

Disley, A. C., D. M. Howard, and A. D. Hunt (2006). "Timbral description of musical instruments". In: *International Conference on Music Perception and Cognition*, pp. 61–68.

Donnadieu, S. (2007). "Mental representation of the timbre of complex sounds". In: *Analysis, synthesis, and perception of musical sounds*. Springer, pp. 272–319.

Elliott, T. M., L. S. Hamilton, and F. E. Theunissen (2013). "Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones". In: *The Journal of the Acoustical Society of America* 133.1, pp. 389–404.

Elliott, T. M. and F. E. Theunissen (2009). "The modulation transfer function for speech intelligibility". In: *PLoS comput biol* 5.3.

Eronen, A. (2003). "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs". In: *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on*. Vol. 2. IEEE, pp. 133–136.

Ethington, R. and B. Punch (1994). "SeaWave: A system for musical timbre description". In: *Computer Music Journal*, pp. 30–39.

Faure, A., S. McAdams, and V. Nosulenko (1996). "Verbal correlates of perceptual dimensions of timbre". In: *4th International Conference on Music Perception and Cognition*, pp. 79–84.

Fedkiw, R., J. Stam, and H. W. Jensen (2001). "Visual simulation of smoke". In: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, pp. 15–22.

Fiebrink, R., D. Trueman, and P. R. Cook (2009). "A Meta-Instrument for Interactive, On-the-Fly Machine Learning." In: *NIME*, pp. 280–285.

Fonteles, J. H., M. A. F. Rodrigues, and V. E. D. Basso (2013). "Creating and evaluating a particle system for music visualization". In: *Journal of Visual Languages & Computing* 24.6, pp. 472–482.

Forbes, A. G., T. Höllerer, and G. Legrady (2013). "Generative fluid profiles for interactive media arts projects". In: *Proceedings of the Symposium on Computational Aesthetics*. ACM, pp. 37–43.

Giannakis, K. (2001). "Sound mosaics: a graphical user interface for sound synthesis based on audio-visual associations." PhD thesis. Middlesex University.

— (2006). "A comparative evaluation of auditory-visual mappings for sound visualisation". In: *Organised Sound* 11.03, pp. 297–307.

Giannakis, K. and M. Smith (2000). "Auditory-visual associations for music compositional processes: A Survey". In: *Proceedings of International Computer Music Conference ICMC2000, Berlin, Germany*. Citeseer.

Giorni, F. and M. Ligabue (1998). "Evangelisti's composition Incontri di Fasce Sonore at WDR: Aesthesic-cognitive analysis in theory and practice". In: *Journal of New Music Research* 27.1-2, pp. 120–145.

Gohlke, K. et al. (2010). "Track displays in DAW software: beyond waveform views". In: *Audio Engineering Society Convention 128*. Audio Engineering Society.

Gounaropoulos, A. and C. Johnson (2006). "Synthesising timbres and timbre-changes from adjectives/adverbs". In: *Applications of Evolutionary Computing*. Springer, pp. 664–675.

Grey, J. M. (1975). *An exploration of musical timbre*. 2. Dept. of Music, Stanford University.

Grey, J. M. and J. W. Gordon (1978). "Perceptual effects of spectral modifications on musical timbres". In: *The Journal of the Acoustical Society of America* 63.5, pp. 1493–1500.

Grill, T. and A. Flexer (2012). "Visualization of perceptual qualities in textural sounds". In: *ICMC*.

Grill, T., A. Flexer, and S. Cunningham (2011). "Identification of perceptual qualities in textural sounds using the repertory grid method". In: *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*. ACM, pp. 67–74.

Heittola, T., A. Klapuri, and T. Virtanen (2009). "Musical instrument recognition in polyphonic audio using source-filter model for sound separation." In: *ISMIR*, pp. 327–332.

Helmholtz, H. L. and A. J. Ellis (1875). *The sensations of tone: As a physiological basis for the theory of music*. Longmans, Green and Co.

Herrera, P. et al. (2000). "Towards instrument segmentation for music content description: a critical review of instrument classification techniques". In: *International symposium on music information retrieval ISMIR*. Vol. 290.

Hsu, W. T. (2011). "On Movement, Structure and Abstraction in Generative Audiovisual Improvisation." In: *NIME*, pp. 417–420.

Johnston, A. (2013). "Fluid Simulation as Full Body Audio-Visual Instrument." In: *NIME*, pp. 132–135.

Jordà, S. et al. (2007). "The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces". In: *Proceedings of the 1st international conference on Tangible and embedded interaction*. ACM, pp. 139–146.

Kendall, R. A. and E. C. Carterette (1993). "Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives". In: *Music Perception*, pp. 445–467.

Kendall, R. A., E. C. Carterette, and J. M. Hajda (1999). "Perceptual and acoustical features of natural and synthetic orchestral instrument tones". In: *Music Perception*, pp. 327–363.

Kohler, W. (1929). "Gestalt Psychology (1929)". In: *Liveright, New York*, p. 34.

Krimphoff, J. et al. (1994). "Characterization of the timbre of complex sounds. 2. Acoustic analysis and psychophysical quantification". In: *J. de Physique* 4, pp. 625–628.

Krumhansl, C. L. (1989). "Why is musical timbre so hard to understand". In: *Structure and perception of electroacoustic sound and music* 9, pp. 43–53.

Krumhansl, C. L. and P. Iverson (1992). "Perceptual interactions between musical pitch and timbre." In: *Journal of Experimental Psychology: Human Perception and Performance* 18.3, p. 739.

Kruskal, J. B. (1964a). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1, pp. 1–27.

— (1964b). "Nonmetric multidimensional scaling: a numerical method". In: *Psychometrika* 29.2, pp. 115–129.

Lipscomb, S. D. and E. M. Kim (2004). "Perceived match between visual parameters and auditory correlates: an experimental multimedia investigation". In: *Proceedings of the 8th International Conference on Music Perception and Cognition*, pp. 72–75.

Lounsbery, M., T. D. DeRose, and J. Warren (1997). "Multiresolution analysis for surfaces of arbitrary topological type". In: *ACM Transactions on Graphics (TOG)* 16.1, pp. 34–73.

Malt, M. and E. Jourdan (2008). "Zsa. Descriptors: a library for real-time descriptors analysis". In: *Sound and Music Computing, Berlin, Germany.*

— (2011). "Real-Time Uses of Low Level Sound Descriptors as Event Detection Functions". In: *Journal of New Music Research* 40.3, pp. 217–223.

Marozeau, J. and A. de Cheveigné (2007). "The effect of fundamental frequency on the brightness dimension of timbre". In: *The Journal of the Acoustical Society of America* 121.1, pp. 383–387.

Marozeau, J., A. de Cheveigné, et al. (2003). "The dependency of timbre on fundamental frequency". In: *The Journal of the Acoustical Society of America* 114.5, pp. 2946–2957.

Mathews, M. V. and J. R. Pierce (1980). "Harmony and nonharmonic partials". In: *The Journal of the Acoustical Society of America* 68.5, pp. 1252–1257.

McAdams, S. (1982). "Spectral fusion and the creation of auditory images". In: *Music, mind, and brain*. Springer, pp. 279–298.

— (2012). "2 Musical Timbre Perception". In: *The psychology of music*, p. 35.

McAdams, S. and J.-C. Cunible (1992). "Perception of timbral analogies". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 336.1278, pp. 383–389.

McAdams, S., S. Winsberg, et al. (1995). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes". In: *Psychological research* 58.3, pp. 177–192.

Melara, R. D. and L. E. Marks (1990). "Interaction among auditory dimensions: Timbre, pitch, and loudness". In: *Perception & Psychophysics* 48.2, pp. 169–178.

Momeni, A. and C. Henry (2006). "Dynamic independent mapping layers for concurrent control of audio and video synthesis". In: *Computer Music Journal* 30.1, pp. 49–66.

Moore, B. C., B. R. Glasberg, and G. M. Proctor (1992). "Accuracy of pitch matching for pure tones and for complex tones with overlapping or nonoverlapping harmonics". In: *The Journal of the Acoustical Society of America* 91.6, pp. 3443–3450.

Müller, M., D. Charypar, and M. Gross (2003). "Particle-based fluid simulation for interactive applications". In: *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, pp. 154–159.

Norman, G. (2010). "Likert scales, levels of measurement and the "laws" of statistics". In: *Advances in health sciences education* 15.5, pp. 625–632.

Nykänen, A. and Ö. Johansson (2003). *Development of a language for specifying saxophone timbre.*

Parise, C. V. and C. Spence (2012). "Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test". In: *Experimental Brain Research* 220.3-4, pp. 319–333.

Patterson, R. D., M. H. Allerhand, and C. Giguere (1995). "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform". In: *The Journal of the Acoustical Society of America* 98.4, pp. 1890–1894.

Patton, K. (2007). "Morphological notation for interactive electroacoustic music". In: *Organised Sound* 12.02, pp. 123–128.

Peeters, G. et al. (2011). "The timbre toolbox: Extracting audio descriptors from musical signals". In: *The Journal of the Acoustical Society of America* 130.5, pp. 2902–2916.

Pitt, M. A. (1994). "Perception of pitch and timbre by musically trained and untrained listeners." In: *Journal of experimental psychology: human perception and performance* 20.5, p. 976.

Platt, J. R. and R. J. Racine (1985). "Effect of frequency, timbre, experience, and feedback on musical tuning skills". In: *Perception & Psychophysics* 38.6, pp. 543–553.

Pratt, R. and P. Doak (1976). "A subjective rating scale for timbre". In: *Journal of Sound and Vibration* 45.3, pp. 317–328.

Prem, D. and R. Parncutt (2007). "The timbre vocabulary of professional female jazz vocalists". In: *Proceedings of International Symposium of Performance Science*, pp. 347–352.

Putnam, L. (2014). "A Method of Timbre-Shape Synthesis Based On Summation of Spherical Curves". In: *Reason* 2, p. 8.

Rasch, R. A. (1978). "The perception of simultaneous notes such as in polyphonic music". In: *Acta Acustica united with Acustica* 40.1, pp. 21–33.

Reeves, W. T. (1983). "Particle systems—a technique for modeling a class of fuzzy objects". In: *ACM Transactions on Graphics (TOG)* 2.2, pp. 91–108.

Renaud, A., C. Charbonnier, and S. Chagué (2014). "3DinMotion A Mocap Based Interface for Real Time Visualisation and Sonification of Multi-User Interactions." In: *NIME*, pp. 495–496.

Rice, S. V. (2005). "Frequency-based coloring of the waveform display to facilitate audio editing and retrieval". In: *Audio Engineering Society Convention 119*. Audio Engineering Society.

Roads, C. (1996). *The computer music tutorial*. MIT press.

Robinson, K. and R. D. Patterson (1995). "The stimulus duration required to identify vowels, their octave, and their pitch chroma". In: *The Journal of the Acoustical Society of America* 98.4, pp. 1858–1865.

Saldanha, E. and J. F. Corso (1964). "Timbre cues and the identification of musical instruments". In: *The Journal of the Acoustical Society of America* 36.11, pp. 2021–2026.

Schönberg, A. (1922). *Harmonielehre*. 3370. Universal-edition.

Schubert, E. and J. Wolfe (2006). "Does timbral brightness scale with frequency and spectral centroid?" In: *Acta Acustica united with Acustica* 92.5, pp. 820–825.

Schubert, E., J. Wolfe, and A. Tarnopolsky (2004). "Spectral centroid and timbre in complex, multiple instrumental textures". In: *Proc. 8th Int. Conf. on Music Perception & Cognition (ICMPC), Evanston*.

Schwarz, D. (2007). "Corpus-based concatenative synthesis". In: *Signal Processing Magazine, IEEE* 24.2, pp. 92–104.

— (2012). "The sound space as musical instrument: Playing corpus-based concatenative synthesis". In: *New Interfaces for Musical Expression (NIME)*, pp. 250–253.

Seago, A. (2009). "A new user interface for musical timbre design". PhD thesis. The Open University.

— (2013). "A new interaction strategy for musical timbre design". In: *Music and human-computer interaction*. Springer, pp. 153–169.

Seago, A., S. Holland, and P. Mulholland (2004). "A critical analysis of synthesizer user interfaces for timbre". In: *In Proceedings of the XVIIIth British HCI Group Annual Conference*. Citeseer.

Shepard, R. N. (1962a). "The analysis of proximities: Multidimensional scaling with an unknown distance function. I." In: *Psychometrika* 27.2, pp. 125–140.

— (1962b). "The analysis of proximities: Multidimensional scaling with an unknown distance function. II". In: *Psychometrika* 27.3, pp. 219–246.

Singh, N. C. and F. E. Theunissen (2003). "Modulation spectra of natural sounds and ethological theories of auditory processing". In: *The Journal of the Acoustical Society of America* 114.6, pp. 3394–3411.

Singh, P. G. and I. J. Hirsh (1992). "Influence of spectral locus and F0 changes on the pitch and timbre of complex tones". In: *The Journal of the Acoustical Society of America* 92.5, pp. 2650–2661.

Smalley, D. (1994). "Defining timbre—refining timbre". In: *Contemporary Music Review* 10.2, pp. 35–48.

— (1997). "Spectromorphology: explaining sound-shapes". In: *Organised sound* 2.02, pp. 107–126.

Spence, C. (2011). "Crossmodal correspondences: A tutorial review". In: *Attention, Perception, & Psychophysics* 73.4, pp. 971–995.

Stam, J. (2003). "Real-time fluid dynamics for games". In: *Proceedings of the game developer conference*. Vol. 18, p. 25.

Stark, A. M. (2014). "Sound Analyser: A Plug-In for Real-Time Audio Analysis in Live Performances and Installations." In: *NIME*, pp. 183–186.

Stevenson, I. (2014). "EMS 2014: The Twelfth Electroacoustic Music Studies Network Conference". In: *Computer Music Journal*.

Takagi, H. (2001). "Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation". In: *Proceedings of the IEEE* 89.9, pp. 1275–1296.

Tervaniemi, M., I. Winkler, et al. (1997). "Pre-attentive categorization of sounds by timbre as revealed by event-related potentials". In: *NeuroReport* 8.11, pp. 2571–2574.

Thoresen, L. and A. Hedman (2007). "Spectromorphological analysis of sound objects: an adaptation of Pierre Schaeffer's typomorphology". In: *Organised Sound* 12.02, pp. 129–141.

Tsai, W.-H. and H.-M. Wang (2006). "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.1, pp. 330–341.

Tzanetakis, G. and P. R. Cook (2000). "Audio information retrieval (AIR) tools". In: *Proc. International Symposium on Music Information Retrieval*.

Tzanetakis, G., G. Essl, and P. Cook (2001). "Audio analysis using the discrete wavelet transform". In: *Proc. Conf. in Acoustics and Music Theory Applications*.

Wapnick, J. and P. Freeman (1980). "Effects of dark-bright timbral variation on the perception of flatness and sharpness". In: *Journal of Research in Music Education* 28.3, pp. 176–184.

Warrier, C. M. and R. J. Zatorre (2002). "Influence of tonal context and timbral variation on perception of pitch". In: *Perception & psychophysics* 64.2, pp. 198–207.

Wedin, L. and G. Goude (1972). "Dimension analysis of the perception of instrumental timbre". In: *Scandinavian Journal of Psychology* 13.1, pp. 228–240.

Wessel, D. L. (1978). "Low dimensional control of musical timbre". In: *Audio Engineering Society Convention 59*. Audio Engineering Society.

— (1979). "Timbre space as a musical control structure". In: *Computer music journal*, pp. 45–52.

Winsberg, S. and J. D. Carroll (1989). "A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model". In: *Psychometrika* 54.2, pp. 217–229.

Winsberg, S. and G. De Soete (1993). "A latent class approach to fitting the weighted Euclidean model, CLASCAL". In: *Psychometrika* 58.2, pp. 315–330.

Wold, E. et al. (1996). "Content-based classification, search, and retrieval of audio". In: *MultiMedia, IEEE* 3.3, pp. 27–36.

Wolpert, R. S. (1990). "Recognition of melody, harmonic accompaniment, and instrumentation: Musicians vs. nonmusicians". In: *Music Perception*, pp. 95–105.

Wu, B., A. Horner, and C. Lee (2014). "Musical Timbre and Emotion: The Identification of Salient Timbral Features in Sustained Musical Instrument Tones Equalized in Attack Time and Spectral Centroid". In: *Proc. 40th Int. Comp. Music Conf.(ICMC)*, pp. 928–934.

Zacharakis, A. (2013). "Musical timbre: bridging perception with semantics". PhD thesis. Queen Mary University of London.

Zacharakis, A., K. Pastiadis, G. Papadelis, et al. (2011). "An Investigation of Musical Timbre: Uncovering Salient Semantic Descriptors and Perceptual Dimensions." In: *ISMIR*, pp. 807–812.

Zacharakis, A., K. Pastiadis, and J. D. Reiss (2014). "An interlanguage study of musical timbre semantic dimensions and their acoustic correlates". In: *Music Perception: An Interdisciplinary Journal* 31.4, pp. 339–358.

Zacharakis, A., K. Pastiadis, J. D. Reiss, and G. Papadelis (2012). "Analysis of musical timbre semantics through metric and non-metric data reduction techniques". In: *Proceedings of the*

12th International Conference on Music Perception and Cognition (ICMPC12) and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM 08), pp. 1177–1182.

Appendices

Appendix A

Ron Arad's Curtain Call Performance: Survey Results

A.1 Audience Questionnaire

Table A.1: Audience Questionnaire Questions 1 - 3

Options	**Total**	**%**
1. What is your gender?		
Female	7	50%
Male	7	50%
2. What is your age group?		
Under 12	1	7.1%
12 - 18	0	
18 - 28	5	35.7%
29 - 39	4	28.6%
40 - 50	1	7.1%
51 - 65	3	21.4%
65 and over	0	
3. What is your professional background?		
Media	3	21.4%
Arts	5	35.7%
Education	2	14.3%
Public Services	1	7.1%
Retail	0	
Entertainment	1	7.1%
Other	2	14.3%

Table A.2: Audience Questionnaire Questions 4 - 7

Options	**Total**	%
4. The visualisations were directly linked to the audio.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	0	
Neither Agree nor Disagree	0	
Somewhat Agree	5	35.7%
Agree	5	35.7%
Strongly Agree	4	28.6%
5. The visualisations responded directly to the audio.		
Strongly Disagree		
Disagree		
Somewhat Disagree	1	7.1%
Neither Agree nor Disagree		
Somewhat Agree	4	28.6%
Agree	5	35.7%
Strongly Agree	4	28.6%
6. Different types of sounds produced different visual qualities.		
Strongly Disagree		
Disagree		
Somewhat Disagree		
Neither Agree nor Disagree	1	7.1%
Somewhat Agree	5	35.7%
Agree	4	28.6%
Strongly Agree	4	28.6%
7. The visualisations and the sounds were entirely separate.		
Strongly Disagree	4	28.6%
Disagree	6	42.9%
Somewhat Disagree	2	14.3%
Neither Agree nor Disagree	1	7.1%
Somewhat Agree	1	7.1%
Agree		
Strongly Agree		

Table A.3: Audience Questionnaire Question 8

Options	**Total**	%
8. The sound played an important element in the performance.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	0	
Neither Agree nor Disagree	0	
Somewhat Agree	0	
Agree	3	21.4%
Strongly Agree	11	78.6%

Table A.4: Audience Questionnaire Questions 9 - 12

Options	**Total**	**%**
9. There were understandable connections between sounds and visuals.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	0	
Neither Agree nor Disagree	1	7.1%
Somewhat Agree	4	28.6%
Agree	7	50%
Strongly Agree	2	14.3%
10. The visual qualities were descriptive of the sound qualities.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	1	7.1%
Neither Agree nor Disagree	3	21.4%
Somewhat Agree	8	57.1%
Agree	2	14.3%
Strongly Agree	0	
11. The visualisations gave meaning to the sounds.		
Strongly Disagree	0	
Disagree	1	7.1%
Somewhat Disagree	1	7.1%
Neither Agree nor Disagree	4	28.6%
Somewhat Agree	4	28.6%
Agree	3	21.4%
Strongly Agree	1	7.1%
12. The visual qualities had no connection to the sound qualities.		
Strongly Disagree	3	21.4%
Disagree	7	50%
Somewhat Disagree	2	14.3%
Neither Agree nor Disagree	1	7.1%
Somewhat Agree	1	7.1%
Agree	0	
Strongly Agree	0	

Table A.5: Audience Questionnaire Questions 13 - 16

Options	**Total**	**%**
13. I could tell when the sounds I made were changing the visualisations		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	1	7.1%
Neither Agree nor Disagree	2	14.3%
Somewhat Agree	2	14.3%
Agree	5	35.7%
Strongly Agree	1	7.1%
Not Applicable	3	21.4%
14. I understood the different ways in which the sounds I made changed the visualisations.		
Strongly Disagree	0	
Disagree	3	21.4%
Somewhat Disagree	0	
Neither Agree nor Disagree	1	7.1%
Somewhat Agree	3	21.4%
Agree	2	14.3%
Strongly Agree	2	14.3%
Not Applicable	3	21.4%
15. I felt I had control over the visual responses through the different sounds I made.		
Strongly Disagree	0	
Disagree	1	7.1%
Somewhat Disagree	3	21.4%
Neither Agree nor Disagree	2	14.3%
Somewhat Agree	2	14.3%
Agree	2	14.3%
Strongly Agree	1	7.1%
Not Applicable	3	14.3%
16. I felt no connection between the sounds I made and the visualisations.		
Strongly Disagree	5	35.7%
Disagree	2	14.3%
Somewhat Disagree	3	14.3%
Neither Agree nor Disagree	1	7.1%
Somewhat Agree	0	
Agree	0	
Strongly Agree	0	
Not Applicable	3	14.3%

Table A.6: Audience Questionnaire Questions 17 - 22

Options	**Total**	**%**
17. I would like to see KIMA again in a different context.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	0	
Neither Agree nor Disagree	0	
Somewhat Agree	2	14.3%
Agree	3	21.4%
Strongly Agree	9	64.3%
18. The musical composition was sonically interesting.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	2	14.3%
Neither Agree nor Disagree	0	
Somewhat Agree	1	7.1%
Agree	4	28.6%
Strongly Agree	6	42.9%
Not Applicable	1	7.1%
19. Did you perceive KIMA more as a performance or an installation?		
Performance	8	57.1%
Installation	5	35.7%
Neither	0	
Other	3	21.4%
20. Have you seen any comparable audio-visual art before?		
Yes	7	50%
No	6	42.9%
Other	1	7.1%
21. Have you been to an Analema Group performance before?		
Yes	7	50%
No	7	50%
22. Would you like to see another Analema Group performance?		
Yes	13	92.9%
No	0	
Other	1	7.1%

A.2 Performers Questionnaire

Table A.7: Performers Questionnaire Question 1

Options	**Total**	**%**
1. Have you been involved in any comparable audio-visual performances, events or installations?		
Yes	3	50%
No	3	50%

Table A.8: Performers Questionnaire Questions 2 - 5

Options	**Total**	**%**
2. There were understandable connections between sounds and visuals.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	0	
Neither Agree nor Disagree	0	
Somewhat Agree	3	50%
Agree	3	50%
Strongly Agree	0	
3. The visual qualities were descriptive of the sound qualities.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	1	16.7%
Neither Agree nor Disagree	2	33.3%
Somewhat Agree	2	33.3%
Agree	1	16.7%
Strongly Agree	0	
4. The visualisations gave meaning to the sounds.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	1	16.7%
Neither Agree nor Disagree	1	16.7%
Somewhat Agree	3	50%
Agree	1	16.7%
Strongly Agree	0	
5. The visualisations had no connection to the sound qualities.		
Strongly Disagree	1	16.7%
Disagree	2	33.3%
Somewhat Disagree	2	33.3%
Neither Agree nor Disagree	1	16.7%
Somewhat Agree	0	
Agree	0	
Strongly Agree	0	

Table A.9: Performers Questionnaire Questions 6 - 9

Options	**Total**	**%**
6. I could tell when the sounds I made were changing the visualisations.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	0	
Neither Agree nor Disagree	1	16.7%
Somewhat Agree	2	33.3%
Agree	3	50%
Strongly Agree	0	
7. I understood the different ways in which the sounds I made changed the visualisations.		
Strongly Disagree	0	
Disagree	2	33.3%
Somewhat Disagree	1	16.7%
Neither Agree nor Disagree	1	16.7%
Somewhat Agree	1	16.7%
Agree	1	16.7%
Strongly Agree	0	
8. I felt I had control over the visual responses through the different sounds I made.		
Strongly Disagree	0	
Disagree	2	33.3%
Somewhat Disagree	0	
Neither Agree nor Disagree	0	
Somewhat Agree	4	66.7%
Agree	0	
Strongly Agree	0	
9. I felt no connection between the sounds I made and the visualisations		
Strongly Disagree	0	
Disagree	5	83.3%
Somewhat Disagree	1	16.7%
Neither Agree nor Disagree	0	
Somewhat Agree	0	
Agree	0	
Strongly Agree	0	

Table A.10: Performers Questionnaire Questions 10 - 13

Options	**Total**	**%**
10. The variations in my vocal performance were accurately represented in the visualisations.		
Strongly Disagree	0	
Disagree	1	16.7%
Somewhat Disagree	2	33.3%
Neither Agree nor Disagree	1	16.7%
Somewhat Agree	2	33.3%
Agree	0	
Strongly Agree	0	
11. The visualisations influenced my vocal performance.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	1	16.7%
Neither Agree nor Disagree	1	16.7%
Somewhat Agree	2	33.3%
Agree	1	16.7%
Strongly Agree	1	16.7%
12. The visualisations were an important factor in my performance.		
Strongly Disagree	0	
Disagree	0	
Somewhat Disagree	0	
Neither Agree nor Disagree	2	33.3%
Somewhat Agree	1	16.7%
Agree	2	33.3%
Strongly Agree	1	16.7%
13. My vocal performance WOULD have been the same without the presence of any visualisation.		
Strongly Disagree	0	
Disagree	3	50%
Somewhat Disagree	0	
Neither Agree nor Disagree	2	33.3%
Somewhat Agree	0	
Agree	1	16.7%
Strongly Agree	0	

Table A.11: Performers Questionnaire Question 14

Options	**Total**	**%**
14. I would be interested in exploring further the use of such such visualisation systems for live performance.		
Yes	6	100%
No	0	