Digital Media Technology Lab
Faculty of Computing, Engineering and the Built Environment
Birmingham City University

Audio Equalisation using Natural Language

Spyridon Stasis

A thesis submitted in partial fulfilment of the requirement for the degree of Doctor of Philosophy, July 2018

Abstract

Equalisation allows a user to control a series of frequency-dependent gains by adjusting the parameters of a network of filters, capable of manipulating the timbre of a sound. In sound production, engineers often use natural language to refer to these timbral transformations, resulting in a shared vocabulary of descriptive terms. This lexicon of semantic terminology allows for simplified and compact descriptions of complex processing actions performed by sound engineers. However, due to the use of natural language to outline these operations, the meaning of the descriptive terms used may be misunderstood, or may possess divergent meanings for different individuals.

The problems inherent to natural language can be alleviated by performing an analysis on the semantic terms that are used in music production and developing computational models based on their function. To perform this analysis crowdsourcing techniques are implemented in order to gather an extensive dataset of terms. In this manner it is possible to exposes the ways in which producers and engineers approach creative audio processing, and this analysis can then be used as a foundation for intuitive interface design. This thesis presents findings from a number of studies on the use of the semantic terminology used in music production, and formalises taxonomies of descriptive terms to provide novel methods for users to interface with equalisation parameters.

Initially, the salience of equalisation in the context of a full processing chain is evaluated. Furthermore, the relationship between a number of key datasets in the field is explored, and synonymous and antonymous definitions within a core list of adjectives is established. In addition, the agreement between descriptive term definitions and the structural similarity of the datasets is analysed. Moreover, the extent to which a term can have multiple definitions, each of which is perceptually divergent, is examined. By clustering different definitions of the same term the concept of semantic sub-representations is introduced. In order to perform this analysis a model of stacked autoencoders is implemented.

The model of stacked autoencoders is then used to create a novel audio production interface, by which users are able to control equalisation parameters based on descriptive language. An unweighted model is first presented, which allows users to navigate between different descriptors using a low-dimensional slider. In turn, signal processing techniques

are implemented in order to make these term definitions adapt to a user's input signal, and can be trained arbitrarily on parameter data.

Overall, this thesis validates the use of descriptive language as a medium for controlling equalisation parameters. It is shown that within this vocabulary, there are consistent relationships between recognised terms, from which a thesaurus of synonymous terminology is constructed. Using these concepts, methods for reducing barriers for inexperienced users are introduced through the development of intuitive abstract interfaces.

Acknowledgements

Firstly, I would like to thank my supervisors Ryan Stables, Jason Hockman and Cham Athwal for their guidance and constant support through my PhD. This research could not have been completed without your invaluable help, insight and meticulous attention.

Also, I would like to acknowledge my past and present colleagues at Birmingham City University: Sean Enderby, Nick Jillings, Matt Cheshire, Sam Smith, Muadh Al-Kalbani, Alan Dolhasz, Xueyang Wang, Carl Southall, Ian Williams, Greg Hough, Dominic Ward, Leo Wang, Tychonas Michailidis, Maciek Tomczak, Jamie Bullock and Salameh Abu Rmeileh. You have all helped me tremendously through the last four years, and you have pushed me to improve and become a better researcher.

I would also like to thank all the Sound Engineering, Music Technology and Sound for Film students at Birmingham City University I had the fortune to teach. I learned a lot through these teaching sessions. And I would also like to thank all of you that participated in my listening tests. Your participation was essential in completing this project.

I would also like to thank the Onassis Public Benefit Foundation for funding my PhD research and specifically my scholarship supervisor Katerina Magkel. This dissertation would not have been possible without your generosity.

Finally and most importantly I would like to thank my family, Georgia, Mina and Denis. No words can describe my gratitude.

Contents

List of Figures

List of Tables

The limits of my language are the limits of my mind. All I know is what I have words for.

— Ludwig Wittgenstein
Tractatus Logico-Philosophicus (1922)

Chapter 1

Introduction

1.1 Background

Sound production is a broad, technical field, where audio is manipulated and prepared to be released as consumer media. The discipline encompasses a number of areas, mainly recording, mixing and mastering for various media, such as music, film and television. As with any other technological discipline, audio production has evolved through the years. Read (1952) provides an overview of the first three eras of sound recording. These were the acoustical, electrical and magnetic eras. However, since 1975 and the digital revolution, sound recording and audio production as a whole have been moving towards the digital domain. As a result, the tools at the sound engineer's disposal have become ever more powerful and versatile.

In audio production, a sound engineer needs to perform a series of complex tasks in order to reach a desired result. Instrument recording relies on the selection of microphones, which have to be appropriate for the chosen source, as well as placement of the microphones in the given space. Mixing these instruments requires a number of parameters to be optimised in order to set the appropriate levels for each individual track (known as balancing), and individual input signals need to be processed using audio effects to creatively shape the final output. Mastering is an extension of the mixing process, by which processing is applied to the master track to prepare it for release.

Whilst recording devices tend to be analog, there have been drastic changes to the manner in which sound engineers mix. The process is now predominantly performed "in

the box", where most sound engineers today take advantage of the ease and versatility that digital audio workstations and digital audio effects offer.

One of the most prominent audio effects is equalisation. Equalisation allows the sound engineer to change the gain of individual frequency bands, either for creative or corrective purposes. According to prominent literature on sound engineering practices, equalisation is the most commonly used tool for altering the timbre of an input sound. The various functions of equalisation, its connection with timbre, and its descriptive terms are discussed in-depth through the remainder of this thesis.

1.2 Motivation

This research is performed to uncover underlying connections between natural language and audio effects. Natural language has the function of providing human-level representations of real-world concepts. In this way, words and phrases allow to simply convey technical ideas and abstract concepts.

However, there is a wide range of differences regarding the agreement and mutual understanding of natural language. It is an intrinsic part of natural language that the words and phrases that are being used may be misunderstood by the target audience, causing the message that is being conveyed to be falsely interpreted. The problem with the definition of abstract concepts is that each person has a different understanding of what these mean and what they entail.

There are many reasons why this failure of communication arises, but the use of words associated with abstract concepts still provides a very powerful tool. Abstract concepts are capable of encapsulating a much more complex structure, that would need a long time to be described. By stating a few simple terms it is possible to convey a very complicated issue, in a way that will be understood by someone familiar with the terminology that is being used.

The analysis of natural language and semantic terminology has been mainly undertaken in the context of timbre research. One of the methods for analysing the timbral qualities of a sound is through the use of semantic terms, a finding that originates from von Helmholtz (1912) analysis of sounds. However, the discipline displays a rigidity with regard to the context of its analysis, remaining to this day focused mostly on musical instruments. Still,

the terms that are used to describe timbre are not used exclusively within the context of musical instruments and can be used to describe other aspects of sound as well. An example is the use of natural language in a sound engineering context. Engineers often colloquially define perceptual characteristics, technical operations, quality and spatial features of sound. Yet, even within the sound engineering and music production disciplines the most common application of natural language comes in the form of adjectives which can be used to describe the timbral quality of a given sound.

These adjectives have also found their way in commercial audio effects in various degrees. Certain equipment, such as the Elyssia xpressor combine low-level technical parameters with settings for altering the timbral quality of the input sound (in the case of the xpressor the *warmth* of the signal). Other manufacturers, such as Waves, have released the OneKnob series of audio effects that completely replaces the technical parameters with a simple dial for controlling a single semantic term. These two methods of using semantic terms in the context of audio effects present two different extremes. On one hand, an on/off setting enabling a timbral transformation is considered a complimentary part of the system and not a central feature, while in the case of the OneKnob series, the effect loses significant versatility in exchange for a simpler interface.

The motivation behind this dissertation is three-fold. Firstly, this study attempts to alleviate the confusion and disagreement that arises from the use of semantic terminology, providing these terms with a more robust form, which will lead to an overall better communication between individuals in the sound engineering discipline. Secondly, this study contributes to the extension of timbre research in a different field. By making use of timbre research practices and methods of analysis, the use of semantic terms in the sound engineering discipline can be illuminated and the study of timbre can be expanded. Thirdly, with regard to the usage of semantic terms in the context of audio effect interfaces, this study presents a middle ground between a complimentary use of these terms and a complete absence of the audio effect's technical parameters. In this manner, the technical aspect of sound engineering can be simplified without losing the versatility provided by the original technical parameters.

1.3 Objectives

The main objective of this work is to assess the nature of natural language in the context of audio effects, more specifically equalisation, and for this reason a series of research questions need to be investigated:

- Is equalisation the primary effect for altering the timbre of an input sound? This needs to be assessed by performing user tests on processing chain decisions, and by investigating the extent to which equalisation is the preferred choice for achieving effects described by timbral adjectives.
- To what extent do people agree on timbral adjectives? As timbre is a very subjective component of sound, this research will need to assess:
  - The agreement between individual definitions of terms
  - The relationship between terms
- Can a single descriptive term encompass multiple discrete representations? By examining the relationship of different definitions of single terms, the extent to which timbral adjectives display homogeneity can be explored. This expands on the common focus on term synonymity in timbre research, introducing an additional layer of complexity in the study of natural language and its perceptual characteristics.
- How can natural language aid music production tasks? Through this research it is possible to uncover the relationship between timbral adjectives and common practices in sound engineering. In this manner, novel systems can be introduced that will form the basis of intuitive tools for music production.

1.4 Methodology

This study relies on a crowdsourcing method for data gathering, similar to the process presented by Cartwright (2016) and Enderby (2017). Previous studies have relied on qualitative methods of data collection, by either using professional mixing engineers to produce

a series of mixes (De Man, 2017), or by interviewing professional sound engineers to better understand their workflow (Pestana, 2013). While the qualitative approach is capable of producing trustworthy results, it is limited by time factors and range of subjects. A quantitative approach, such as crowdsourcing, is capable of providing a larger corpus of data, from a wider range of subjects, but presents the disadvantage of that data being less trustworthy. However, given that there is a data preprocessing stage, outliers are identified and eliminated, these problems can be alleviated. Furthermore, throughout the thesis the results from the crowdsourced dataset are verified through additional experiments, in perceptual listening tests and user testing.

The first objective of this thesis is to assess whether equalisation is the primary processing module for altering the timbral quality of an input sound. In order to evaluate this empirical claim, supported by the sound engineering literature, it is necessary to perform a study regarding the ways in which sound engineers make use of audio effects. It is generally well known that sound engineers make use of processing chains in order to creatively or correctively transform an input sound, and a study on this topic can potentially verify the role of equalisation as the main tool for altering the timbre of a sound.

The second objective of this thesis requires an investigation of the between-term relationships. As a shared, universal vocabulary of semantic terminology does not exist, it is necessary to assess the relationships between terms. In this instance performing an analysis through a single dataset is considered limiting and instead a comparative analysis between three datasets is chosen. This allows for assessing the existence of structural similarities as well as the existence of universal agreement and confidence between terms.

While the second objective investigated the between-term relationships of semantic terms, the third objective is to investigate the within-term relationships. Considering that disagreements arise from divergent uses of the same term, it is imperative to consider the existence of deviations within a single term context. For this reason an analysis of the dataset is performed for isolated terms and perceptual testing is implemented to verify whether a term can take perceptually divergent forms.

Finally, the fourth objective is to incorporate the knowledge gained from the study of semantic terms in novel music production tools. The quantitative dataset in this case provides a very good basis for training machine learning models for that task. Recent advancements in machine learning and deep learning, in the form of recommender systems,

generative models and more traditional clustering methods are adapted to operate within a music production context. The systems are then evaluated through perceptual testing and user testing to assess their performance.

1.5 Thesis structure

Chapter 2 presented a literature review that includes an overview of timbral research in relation to the use of natural language in sound engineering. The chapter further discusses the empirical methods used by sound engineers in order to alter the timbral quality of an input sound, and pinpoints equalisation as the main processing unit where natural language is used to explain timbral transformations.

Chapter 3 presents an analysis of the structure of processing chains in order to semantically adapt an input signal to a given term using multiple audio effects. The study suggests that equalisation is the main processing unit applied in order to alter the timbre of a signal when provided with natural language cues. For this reason it is found necessary to investigate equalisation in isolation, and examine parameter settings for achieving a range of timbral adjectives.

Chapter 4 presents a comparative review using three key datasets containing semantic equalisation terminology. The chapter finds structural similarities between all datasets, but also deviations in the confidence of subjects for definitions of each term. Therefore it is considered that the hierarchy of semantic terminology can include different definitions for a single term. To further investigate this complexity of timbral adjectives a further layer of analysis will be performed, assessing the perceptual characteristics of each term separately.

Chapter 5 examines the existence of sub-categories (sub-representations) of two of the most common terms in a crowd-sourced dataset, *warm* and *bright*. To this end, clustering techniques are applied in order to identify the principal sub-representations of each descriptor, and subsequently perceptual testing is performed in order to investigate the extent to which these sub-representations exhibit perceptual differences.

Chapter 6 draws on the knowledge of the previous chapters in presenting an intuitive system that allows for timbral transformation between two distinct descriptive terms. By performing a comparative review on the performance of dimensionality reduction and reconstruction (regression and interpolation) methods, the technique that provides the most

appropriate mapping in a low-dimensional space and the highest reconstruction accuracy is identified. This system is used to navigate between two timbral adjectives, by taking into account new input characteristics, through an adaptive architecture. Finally, user testing is performed, and it is found that the proposed model is capable of reconstructing the parameters necessary for achieving *warm* and *bright* characteristics.

Chapter 7 summarises the main findings and contributions of the thesis, and proposes further areas of investigation for future research.

1.6 Associated publications

Parts of this thesis have been published in various international conferences and journals. These are detailed below:

- A short summary of Chapter 3 was presented at the 3rd Workshop on Intelligent Music Production (WIMP 2017) (Stasis, Jillings, Enderby and Stables, 2017b), and was extended for publication at the 20th International Conference on Digital Audio Effects (DAFx 2017) (Stasis, Jillings, Enderby and Stables, 2017a).

- An initial, experimental study of Chapter 5 was published at the 2nd Workshop on Intelligent Music Production (WIMP 2016) (Stasis, Hockman and Stables, 2016), which was then expanded and published at the International Conference on New Interfaces for Musical Expression (NIME 2017) (Stasis, Hockman and Stables, 2017).

- The initial research for Chapter 6 was presented at the 18th International Conference on Digital Audio Effects (DAFx 2015) (Stasis et al., 2015) and an extended version was published in a special Audio Signal Processing edition of Applied Sciences (Stasis, Stables and Hockman, 2016).

Chapter 2

Background

2.1 Timbre

Sound is a multidimensional structure encompassing the dimensions of pitch, loudness, duration and timbre. Some researchers even include spatial position as a fundamental component (Zacharakis, 2013). The difficulty in sound research is based on the fact that certain dimensions of sound are clearly defined, while others exhibit diverging perceptual attributes. Even though loudness and pitch exhibit significant perceptual differences between listeners, they can each be measured along a single dimension, frequency (Hz) for pitch and power (dB) for loudness. However, the dimension that has been the most elusive in both its definition and its perceptual quality through the literature is *timbre*.

Researchers have been attempting to deconstruct the dimension of timbre from as early as von Helmholtz (1912), which provided some of the first investigations into the association of the sound dimension with acoustic correlates (audio features) and verbal adjectives, which are also known as timbral adjectives or semantic terms. These acoustic correlates encompass features that are extracted from a sound and measure its temporal, spectral, spectrotemporal or power-related information, while timbral adjectives describe the perceptual attributes of a sound, for instance if it appears to be *bright*, *warm*, *harsh* and so on.

Despite the early work of von Helmholtz (1912) on timbre, the definition of this dimension was produced much later and it is attributed to the ANSI (1973) standard, which defines timbre as *the attribute of auditory sensation in terms of which a listener can judge*

two sounds similarly presented and having the same perceived loudness, pitch and duration as being dissimilar. This viewpoint has become the standard through literature, and many further investigations on the nature of timbre use this definition as a starting point.

Whilst the definitions of pitch, loudness and duration are well understood and agreed upon, reaching a consensus on the nature of timbre presents a difficulty (Jensen, 1999). The biggest problem with the current definition of timbre, as Zacharakis (2013) argues, is that it is a definition by negation, meaning that instead of defining what timbre is, the ANSI (1973) instead defines which aspects of audio should not be used to identify timbre. Through the years there have been many other instances of researchers finding the definition of timbre problematic (Sankiewicz and Budzyński, 2007; Papanikolaou and Pastiadis, 2009; Donnadieu, 2007).

Bregman (1994) suggests that the the root of the problem in defining timbre is found in the definition provided by ANSI (1973). Due to the definition it is not possible to compare two different sounds that do not posses a pitch, for example non-musical, everyday sounds such as a hammer hitting a nail. Alternatively, Krumhansl (1989) makes the suggestion that the problem exists due to the researchers' association of timbre with classical instruments, a practice that could have caused a narrowed view on the nature of timbre. However, the problem most likely arises due to an essential difference of timbre and the remaining auditory dimensions of pitch, loudness and duration. While the other dimensions of sound, as in pitch, duration and loudness exhibit varying levels of agreement regarding their measurement and definition, timbre does not display the same behaviour. Where pitch, loudness and duration can be measured in single units and in a single domain, the frequency being measured in Hz in the frequency domain, the loudness in decibels in the power domain, and the duration in seconds in the time domain, timbre does not operate in a single measurable domain.

From these studies, some researchers have argued replacing the word timbre, due to the disagreements associated with its perception and its problematic definition, to terms such as sound color (Slawson, 1985) or tone color (Erickson, 1975). Other positions however take a more extreme approach, arguing that the term and its definition should be excluded from sound research altogether (Martin, 1999). Even though this extreme position does not provide a solution to the issue, it is an indication towards the problems timbre research has posed and the progress in objectively defining timbre and its resulting semantic terms.

2.1.1 Multidimensional nature of timbre

Rossing et al. (2002) presents the example of describing sounds in terms of pitch with words such as *high* and *low*, with regards to the position of the fundamental frequency in a magnitude spectrum. Similarly for loudness, individuals can use terms such as *quiet* and *loud* in an attempt to compare the power in dB. These perceptual attributes are still subject to more complicated principles, such as masking in the case of loudness measurements, and the intensity of the fundamental frequency in the case of pitch tracking. But it can be established that loudness and pitch, as well as duration, operate along single dimensions and that they are mainly associated with single features. The pitch can be measured by calculating the fundamental frequency, the loudness by measuring the total loudness, as described by Moore et al. (1997), and the duration by measuring the time extent of a signal from its starting point to its end point. The same behaviour is not exhibited in the case of timbre. In contrast, timbre operates in a multidimensional space, a remark made early on by Licklider (1951). Furthermore, not only is timbre defined in a high-dimensional space, but it can also potentially affect the perception of other auditory dimensions (pitch, loudness and duration) (Krumhansl, 1989).

A method of unpacking the dimensions of timbre is through feature extraction, which according to Bullock (2008), aims to reduce the dimensions of audio signals. An unprocessed audio signal offers a noisy input for analysis, with information that can be considered redundant. However, it is possible to derive statistical information regarding the audio signal from the raw audio, known as audio features. Feature extraction can then lead to the establishment of different representations for music analysis, as described by Bello (2016) and presented in Figure 2.1. The low-level representations will include features which correlate with attributes of the input signal, such as the fundamental frequency, which correlates with pitch perception or the RMS amplitude, which correlates with loudness perception. The mid-level representations provide a further layer of information regarding musical characteristics, in rhythm, tempo and harmony. The high-level representations encompass aspects of the input sound regarding the instrument/source, artist, genre, as well as abstract definitions regarding the semantic terminology used to describe a sound (Herrera et al., 2005).

Finally, Rubio (2005) makes a distinction between the low-level and high-level representations, stating that the first are related to the audio signal itself, and are not important to

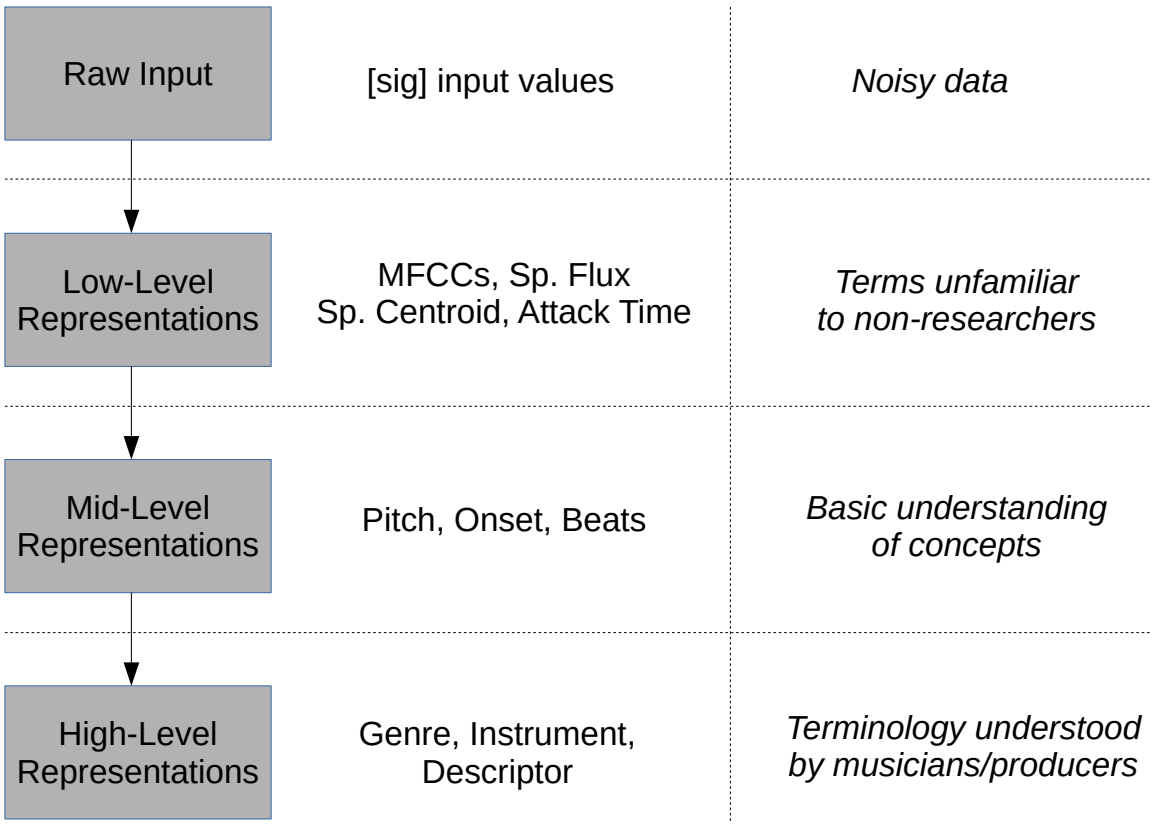


Figure 2.1: Levels of audio representations according to Bello (2016). Here, each layer corresponds to a different set of features/qualities of the input and their respective understanding by non-researchers.

the non-researcher, while the high-level representation possess semantic value and contain an understandable meaning to the average listener. However, it is through the low-level representations that an understanding for the high-level representations can be acquired computationally.

2.1.2 Methods of analysis

For the most part, research on timbre has followed the guideline of the ANSI (1973) standard. To that end, researchers have compared different sounds, which possess the same pitch, loudness and duration. As Hajda (2007) explains, this type of analysis focuses on the problem of classification. The end goal of these studies is to recognise the source of

the sound, in other words the instrument that is being played. Following the classification problem, the next step is to compare different sounds using relational measures (Hajda, 2007). This approach still takes into consideration the ANSI (1973) definition of timbre, and presents a series of audio samples to subjects. The goal of this approach is to find an estimate of the similarity/dissimilarity between audio samples.

Hajda (2007) makes the categorisation of timbre research as either verbal or non-verbal, as is illustrated in Figure 2.2. Both processes, as described below, are based on perceptual grading of different sounds, but differ on their choice of rating methods and the techniques applied to visualise the resulting perceptual spaces.

- **Non-verbal**: This approach encompasses the rating of sounds based on their perceptual, pairwise similarity/dissimilarity, and it has been the more popular method of timbre analysis (Plomp, 1970, 1976; Grey, 1977; Miller and Carterette, 1975; Iverson and Krumhansl, 1993; Caclin et al., 2005). This method implements a technique known as Multidimensional Scaling (MDS) in order to create a timbre space.

- **Verbal**: The second method in measuring the relationship between audio samples is by grading sound in terms of semantic scales, either through the Verbal Attribute Magnitude Estimation (VAME) (Kendall and Carterette, 1993), or the semantic differential (von Bismarck, 1974) method. For this reason, the analysis method implements dimensionality reduction techniques, most commonly Principal Component Analysis (PCA) or Factor Analysis (FA) in order to construct the resulting semantic space.

Timbre space

In the non-verbal analysis method, researchers implement MDS, a method for finding, measuring and visualizing similarities between instances of a dataset, initially proposed by Torgerson (1952). The technique is extremely useful in dealing with datasets operating in a high-dimensional space, being able to provide comparison of the different entries, identifying the most important dimensions of the dataset, and visualizing the instances in a space of reduced dimensionality that humans can perceive.

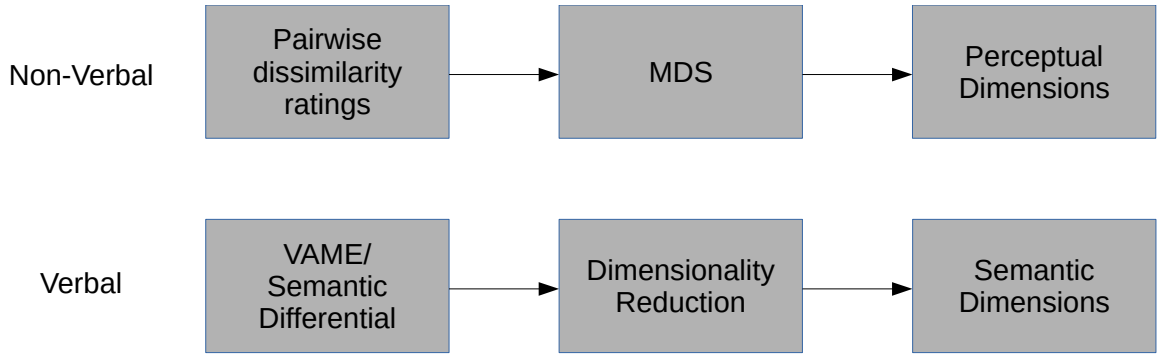


Figure 2.2: The two different methods of timbre analysis and their respective procedures. The verbal approach makes use the semantic differential methods and dimensionality reduction to construct a semantic space, while the non-verbal approach implements pairwise dissimilarity ratings and multidimensional scaling to uncover perceptual dimensions.

Techniques In the classical MDS form, the algorithm takes into account that the similarity/dissimilarity between data entries is measured in Euclidean space, with the distance d defined as in Equation 2.1:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{2.1}$$

where x and y detail the position in a two-dimensional Cartesian space. The algorithm then follows a set number of steps in order to produce the low-dimensional coordinates for n data entries:

1. Set up the matrix of squared distances $P^{(2)} = [d^2_{i,j}]$
2. Apply double centering so that $B = -\frac{1}{2}KP^{(2)}K$, where $K = I - n^{-1}O$ and I is the identity matrix, and O is an $n \times n$ matrix of ones.
3. Extract the m largest eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_m)$ of B and their corresponding eigenvectors $(\epsilon_1, \epsilon_1, ..., \epsilon_m)$
4. Derive the m-dimensional coordinates of the n data points from the coordinate matrix $X = E_m\Lambda_m^{1/2}$, where E_m the matrix of m eigenvectors and Λ_m the diagonal matrix of m eigenvalues of B respectively.

However, as Wickelmaier (2003) suggests, assuming that proximities will behave like distances is not always correct, and for this reason researchers such as Kruskal (1964b) and Shepard (1962) have proposed nonmetric methods of MDS. In this application the MDS algorithm attempts to minimize the squared difference of proximities (p) and distances (d) between data points. The optimisation function is known as stress, and is described in Equation 2.2.

$$Stress = \sqrt{\frac{\sum(f(p) - d)^2}{\sum d^2}} \tag{2.2}$$

In more detail, the steps that need to be performed in order to compute nonmetric MDS are as follows:

1. Assign random coordinates to the datapoints
2. Calculate the distances (d) between the data points
3. Find the optimal monotonic transformation ($f(p)$) of the proximities
4. Minimize the stress (Equation 2.2) between the distances d and the optimally scaled data.
5. Compare the stress to the stress criterion. If the stress is considered acceptable then exit the algorithm, otherwise return to step 2.

MDS in timbre research The first application of MDS in timbre research was performed by Plomp and Steeneken (1969) and in turn more researchers followed this example (Wessel, 1973; Miller and Carterette, 1975). Grey (1977) further expanded the field by performing the perceptual testing using sound synthesis, which allowed for significantly more control over the auditory dimensions of pitch, loudness and duration. In addition, Grey (1977) produced a visualization of the timbre space, projecting the similarity/dissimilarity ratings in a low dimensional mapping (Figure 2.3). Following this approach, numerous studies have implemented the MDS algorithm, in order to visualize the resulting space, but also to acquire the salient perceptual dimensions in the dataset (Kendall and Carterette, 1991; Iverson and Krumhansl, 1993; McAdams et al., 1995).

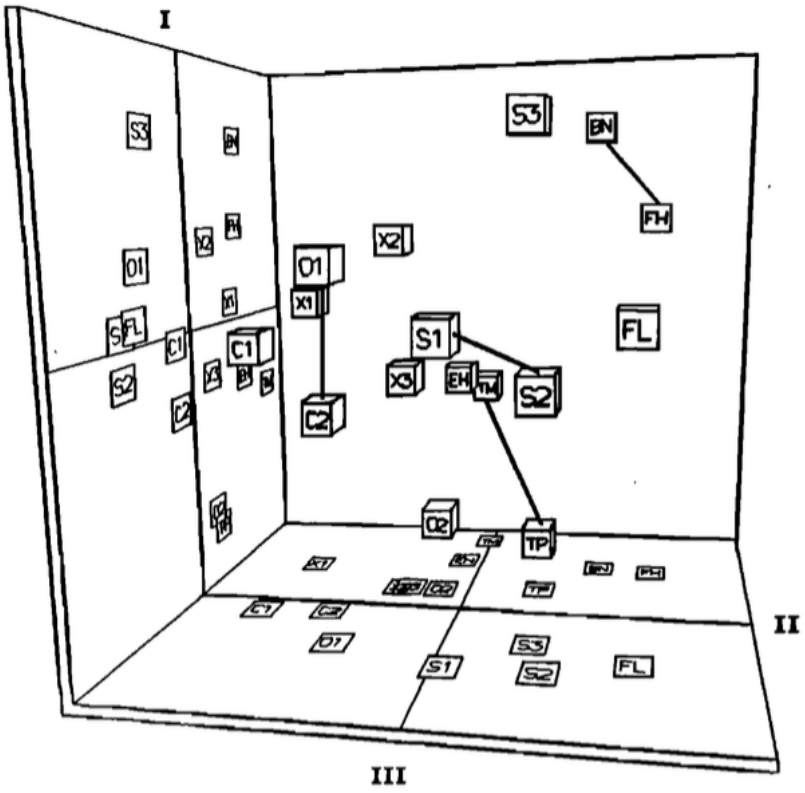


Figure 2.3: An example of the resulting perceptual space using MDS as produced by Grey and Gordon (1978). The points in the space represent different instruments and their corresponding relationship. The correlation between dimensions (I, II, III) and audio features is measured to find the audio features responsible for the structure of the perceptual space.

In the case of perceptual testing regarding audio, MDS poses a robust method for comparing the different audio instances. As explained by McAdams (1999), pairwise similarity testing and MDS display a number of advantages compared to other methods of analysis. The primary advantage is that the task presented to the subjects is simple, asking them to only judge two sounds with respect to their similarity. Additionally, the technique does not force the researcher into making any a priori assumptions regarding the nature of the perceptual structure.

In his in-depth paper on the use of MDS in timbre analysis, McAdams (1999) provides an overview of the methodology that this research approach needs to implement. There is a standard series of steps that researchers undertake in order to construct the timbre

spaces, beginning by acquiring a number of samples from different sources, which need to be normalised, and need to have the same pitch, loudness and duration, as per ANSI (1973) guidelines.

Figure 2.4 presents the process for the non-verbal, MDS approach. Here, the set of sounds is presented in pairs to a group of listeners, with the subjects asked to rate the degree of dissimilarity between each pair. Furthermore, distance matrices are constructed from the responses of each subject, which is represented as a non-symmetric $N \times N$-dimensional matrix. The non-symmetric quality of the matrix can then help in identifying comparisons that the subjects have difficulty in judging. Finally, MDS is applied to the set of matrices, fitting a distance model to the dissimilarity data. The only variables in the process are the choice of audio samples and MDS algorithm that the researcher will use.

Traditionally, once MDS is applied and the salient perceptual dimensions are identified, researchers attempt to uncover correlations between perceptual dimensions and audio features. The magnitude weighted average frequency, known as the spectral centroid, was one of the first audio features to be used as a measure of instrument quality (Knopoff, 1963), and was sequentially found to correlate with one of the perceptual dimensions in studies involving MDS, such as Grey and Gordon (1978). Follow-up studies have attempted to find correlations between all the salient dimensions of timbre studies, which have resulted in opposing and inconsistent results.

Hajda et al. (1997) has delivered the main criticism of the MDS approach, which is the dependency of the similarity measurements on the input stimuli. As different researchers will use different samples, instead of a standardised dataset of samples, there will always exist an inconsistency between listening test results. This criticism is also apparent in the resulting dimensions of importance that have been found in such studies, where the only audio feature that consistently appears through the different tests is the spectral centroid (Fuhrmann, 2012). Other features that tend to be of relevance to the perceptual dimensions in the majority of studies include the attack and decay transients (Lakatos, 2000), the time-varying spectral shape (McAdams et al., 1995) or the spectrum's fine structure (Caclin et al., 2005).

MDS variations This inconsistency between important dimensions has led to different implementations of the MDS method. McAdams (1999) has also provided a cohesive list

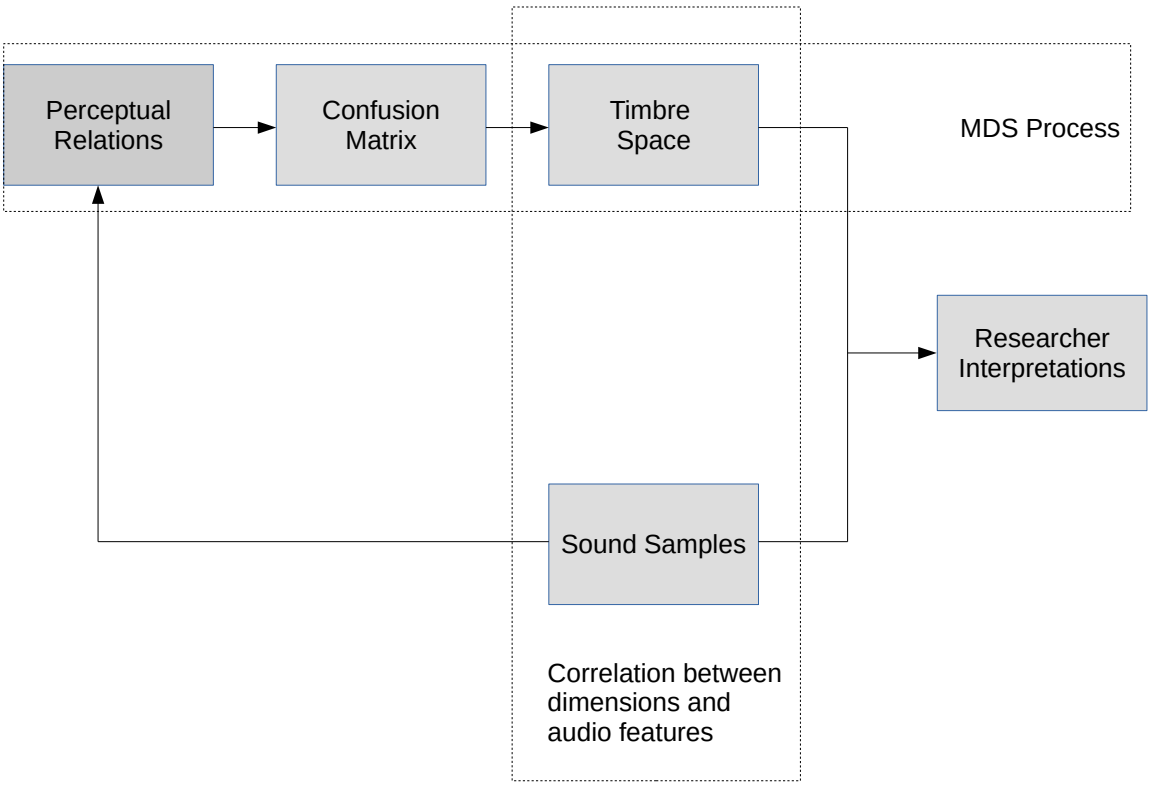


Figure 2.4: Process of perceptual testing and analysis for construction of timbre spaces, as presented by McAdams (1999). The researcher provides the test subjects with audio samples and a confusion matrix is produced through similarity ratings. From this process a timbre space is constructed and the correlation between its dimensions and the original audio samples features to interpret the structure of the perceptual space.

of the different types of MDS modifications that have been implemented in timbre analysis publications, beginning with the use of Euclidean models (as explained in Equation 2.1), and Minkowski generalizations of Euclidean models, defined in Equation 2.3. This formulation was used in the studies of Wessel (1979) and Kruskal (1964a,b) and is known as the MDSCAL method, where x_{ik} is the coordinate of timbre i on dimension k, K the total number of dimensions and r the Minkowski metric variable.

$$d_{ij} = [\sum_{k=1}^{K}(X_{ik} - X_{jk})^r]^{1/r} \tag{2.3}$$

Overall, there are two distinct problems with the use of the classic form and Minkowski

generalisation of MDS:

- the set of dimensions, and their importance, can differ between listeners
- not all input stimuli should be expected to share the same perceptual dimensions

An approach that bypasses the first problem, proposed by Carroll and Chang (1970), is INDSCAL, which has been used in a number of studies, including Miller and Carterette (1975), Grey (1977) and Grey and Gordon (1978). As is displayed in Equation 2.4, where d_{ij} the distance between timbres i and j on the k-th dimension and w_{nk} represents the weighting applied to the k-th dimension of the n-th listener. However, this method also comes with a significant drawback, in that the number of instances in the model will increase each time a new listener is added to the dataset.

$$d_{ij} = [\sum_{k=1}^{K} w_{nk}(X_{ik} - X_{jk})^2]^{1/2} \tag{2.4}$$

A solution to this issue was presented by Winsberg and De Soete (1993), with the introduction of the “latent classes” concept in the CLASCAL method, where it is considered that the subjects will form an n-number of groups. The system then uses statistical methods to allocate each subject to a class. In that manner the individual weights are replaced by the “latent classes” weights. However, another problem arises in regard to the inconsistent characteristics between input stimuli, known as “specificities”. For this reason Equation 2.5 was proposed, where s_i and s_j are "specificities" corresponding to timbres i and j.

$$d_{ij} = [\sum_{k=1}^{K}(X_{ik} - X_{jk})^2 + s_i + s_j]^{1/2} \tag{2.5}$$

An improved system was suggested by Winsberg and Carroll (1989) in their extended EXSCAL model, paving the way for a system proposed by McAdams et al. (1995). The system combines the “latent classes” and “specificities” solutions, applying weights that would explain both differences in the listeners’ perception and the timbral differences of the input stimuli. This is presented in Equation 2.6, where w_{kc} is the weight on the k-th dimension for the c class, and v_c the weight on the set of "specificities".

$$d_{ij} = [\sum_{k=1}^{K}(w_{kc}(X_{ik} - X_{jk})^2 + v_c(s_i + s_j))]^{1/2} \tag{2.6}$$

Semantic space

One limitation of the MDS approach to timbre and the non-verbal approach in general, as Zacharakis (2013) argues, is the inability to evaluate semantic terms within the analysis method. Semantic terms have been an important aspect of timbre since very early research, with von Helmholtz (1912) being among the first to attempt an interpretation of semantic terms with regard to audio characteristics.

To this end, researchers have based their studies on the assumption that a given sound can be adequately described by the use of timbral adjectives (von Bismarck, 1974; Kendall and Carterette, 1993). The first method implemented was the semantic differential, where subjects were asked to rate a sound in the scale of two opposing terms, for example *full* and *empty*. While this method was implemented by Lichte (1941) and von Bismarck (1974), Kendall and Carterette (1993) proposed a variation of this approach, known as VAME, where instead of opposing timbral characteristics, the extremes of the scale where labelled by a descriptor and its negation, as in *full* and *not full*.

The number of dimensions in these studies is defined by the number of scales the subjects were asked to grade a sound on. Therefore, in order to visualize the semantic space the researchers implemented dimensionality reduction techniques, such as PCA and FA, instead of MDS.

Dimensionality reduction Extracted data from sources such as digital photos, video, x-ray scans and audio signals usually display a high-dimensionality. Due to the high number of variables present in such datasets, their visualisation and interpretation can prove difficult. In such cases dimensionality reduction can be implemented in order to scale down the number of dimensions in a low-dimensional representation (Van Der Maaten et al., 2009). This process is depicted in Figure 2.5.

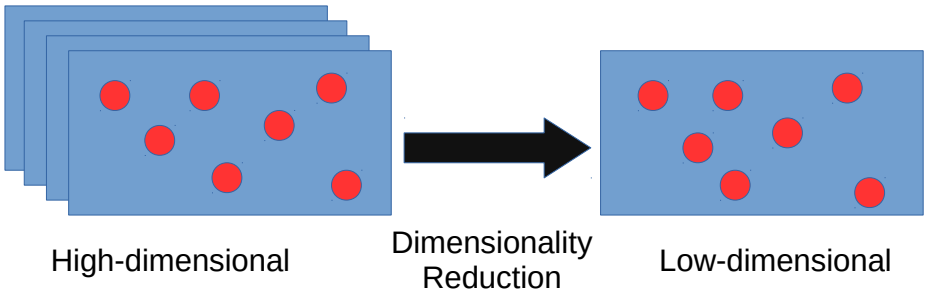


Figure 2.5: The process of dimensionality reduction, by which a high-dimensional space (right) is undergone a transformation in order to produce a low-dimensional representation (left). It is highly desirable that the structure of the original dataset is retained in the low-dimensional mapping.

Dimensionality reduction can provide solutions to a number of research problems. As the technique is reducing the number of parameters while preserving the core information of the original dataset, it essentially applies compression. In that regard, dimensionality reduction can reduce the time needed to train machine learning models or minimise the storage space allocated. A high-dimensional space is very difficult to visualise, and as Van der Maaten and Hinton (2008) suggest, data visualisation is an important problem that dimensionality reduction can solve through the low-dimensional mappings it produces. Finally, Jimenez and Landgrebe (1998) state that dimensionality reduction methods can bypass a number of problems regarding the high-dimensional dataset properties, such as the existence of collinearity (Dormann et al., 2013), and address the curse of dimensionality, by which the time required for optimising a system will grow factorially as the number of dimensions grows.

Many different dimensionality reduction techniques have been proposed, with Van Der Maaten et al. (2009) providing a taxonomy based on their individual characteristics (Figure 2.6). The techniques are split into convex and non-convex groups. Convex techniques are defined as methods capable of optimising an objective function that contains only global minima, while non-convex systems optimize an objective function that contains local minima.

The convex group is further separated into full spectral and sparse spectral techniques. Full spectral techniques will perform an eigendecomposition of a full matrix, which encompasses the covariances between dimensions, or the pairwise distances between data-

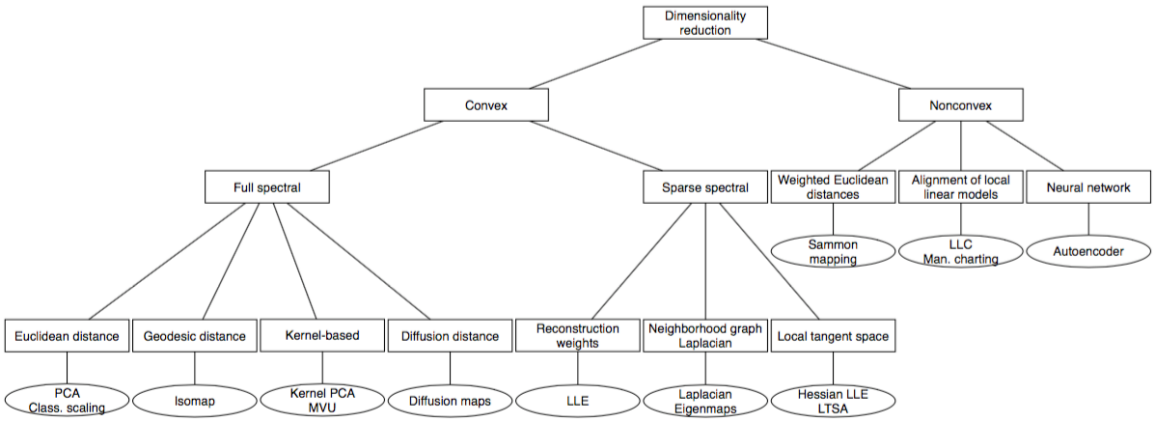


Figure 2.6: Taxonomy of dimensionality reduction techniques, split between convex and non-convex methods, as defined by Van Der Maaten et al. (2009).

points. Different distances will result in different techniques, where Euclidean distance is associated with PCA (Pearson, 1901; Hotelling, 1933) and classic MDS (Torgerson, 1952), geodesic distance is used to perform the Isomap method (Tenenbaum et al., 2000), diffusion distance will produce Diffusion Maps (Lafon and Lee, 2006), and the calculation of the eigenvectors from the kernel matrix instead of the covariance matrix will perform kernel PCA (Schölkopf et al., 1998). On the other hand, sparse spectral techniques solve an eigenproblem for only a portion of the covariance matrix. This can be performed through the reconstruction of weights, a technique known as Local Linear Embedding (LLE) (Roweis and Saul, 2000), calculated through a neighbourhood graph, known as Laplacian Eigenmaps (Belkin and Niyogi, 2002), or calculated through the local tangent space, in techniques such as Hessian LLE (Donoho and Grimes, 2003) and Local Tangent Space Analysis (LTSA) (Zhang and Zha, 2004).

On the other hand, non-convex techniques do not share as many similarities with each other as convex techniques do. Some examples of non-convex dimensionality reduction methods implement weighted Euclidean distances, in the case of Sammon mapping (Sammon, 1969), the alignment of local linear models, in Locally Linear Coordination (LLC) (Roweis et al., 2002), and finally the implementation of Artificial Neural Networks (ANN), such as a system of stacked autoencoders (Hinton and Salakhutdinov, 2006).

Despite the large number of techniques developed in the field, Van Der Maaten et al. (2009) has shown that nonlinear and non-convex techniques for dimensionality reduction

are still incapable of outperforming their linear counterparts, such as PCA.

Semantic space subjectivity problem The semantic space approach for timbre analysis presents some of the same problems described by Hajda (2007), as the MDS-based timbre space methods. Firstly, the dependence of the study on the input stimuli — inherent in any timbre related perceptual test — is not bypassed through the use of semantic terms. Moreover, the semantic space approach introduces an additional problem in the choice of the semantic terms used for rating by the subjects, as well as the language used in the study, with research showing disagreement even between UK and US English speakers (Disley and Howard, 2004).

2.1.3 Audio features

McAdams et al. (1995); McAdams (1999) suggests that timbre is an auditory feature which contains a number of perceptual dimensions. These perceptual dimensions are derived from the similarity/dissimilarity ratings provided by the subjects, and uncovered through the application of MDS. For instance, it is usually found that the first perceptual dimension will illustrate the difference of musical instruments considering the extent to which they are percussive. In this context the nature of the initial transient of the sample influences the perception of the subjects.

Following the identification of the perceptual dimensions, researchers aim to find correlation between these and different audio features. Through the years, a plethora of audio features has been defined, which can aid in identifying and measuring these perceptual dimensions of timbre.

In his analysis of existing studies regarding the perception of timbre, McAdams et al. (2006) states that there are a total of 72 features, which have been found to correlate with perceptual dimensions. These can be split into four categories: temporal, spectral, spectrotemporal and energetic, which also correlate with Peeters et al. (2011) categorisation of representations from which audio features arise. From the four categories the only one that is not associated with timbre is energetic, which includes information regarding the loudness of a signal, and therefore it does not coincide with the ANSI (1973) definition. In addition, researchers have used abstracted audio features for timbre analysis, such as

the Mel-frequency Cepstral Coefficients (MFCC) (De Poli and Prandoni, 1997; Loughran et al., 2008; Brent, 2009).

Temporal

The temporal category includes features which are calculated from time segments of the audio signal (Peeters et al., 2011). These are further separated into two groups, global features calculated from the whole duration, and time-varying features, calculated from segments of the audio. The temporal features have been found to correlate with perceptual dimensions in a wide number of studies (Lakatos, 2000; Iverson and Krumhansl, 1993; Krumhansl, 1989; Grey, 1977; Grey and Gordon, 1978). However, these results have been criticised for their inconsistency in the existing literature with regard to the diverging audio features that have been found to correlate with the different perceptual dimensions. According to Hajda (2007) the problem arises from the lack of robust operational characteristics regarding the audio information, and also due to computational errors produced by some of the feature extraction algorithms (McAdams, 1999). The temporal characteristics include features such as instances of the amplitude envelope (attack and decay times), temporal centroid and zero-crossing rate.

Spectral

In contrast to temporal features, which include both static and time-varying measurements, the spectral features are extracted solely in the frequency domain. This category encompasses statistical measurements of the spectrum, i.e. the spectral centroid, spectral spread, spectral skewness and spectral kurtosis, as well as characteristics regarding the slope of the spectrum in spectral roll-off and spectral slope. Spectral features have shown correlation with the first perceptual dimension in Grey (1977); Grey and Gordon (1978), and the spectral centroid can be considered an indication towards the *brightness* of a signal (Beauchamp, 1982).

Spectral features also include harmonic measurements of the spectrum. These entail metrics such as noisiness, irregularity, tristimulus, inharmonicity and odd to even harmonic ratio.

Spectrotemporal

The spectrotemporal characteristics of a sound measure the variation of the spectral information over time. This generally means features are extracted from a short-term Fourier transform (STFT) using frame-based analysis, as opposed to a long term average spectrum (LTAS). The main feature in this category is spectral flux, also known as spectral variation. Similar to the spectral category, before spectral flux was proposed as a feature by Krimphoff et al. (1994), researchers were describing its presence through different terminology. Grey (1977) found certain variations of the spectral dimension over time, which researchers have called the attack synchronicity (McAdams et al., 2006) to correlate with perceptual dimensions. Similarly, the attack centroid found by Grey (1977) corresponds to the spectral flux, detailing the presence of precedent high frequency, low amplitude energy on the attack partition of the sample, which the researcher suggests takes the characteristic of inharmonic energy.

2.1.4 Dimensions of timbre

Perceptual dimensions

Research on timbre has often focused on finding a correlation between specific audio features and perceptual dimensions. In his comparative work on timbre analyses McAdams et al. (2006) noted that the audio features that frequently correspond to timbral dimensions using the non-verbal method are: spectral centroid, spectral deviation, spectral density, attack time, decay time, amplitude envelope, spectral flux, pitch strength, attack synchronicity, attack centroid and noisiness. In her thesis, Emiroglu (2007) proposed a hypothetical timbre model, which combined audio features found in existing literature to correlate with perceptual dimensions. The shared features with the McAdams et al. (2006) study included: amplitude envelope (Lakatos, 2000), spectral density (Lakatos, 2000), attack centroid (Iverson and Krumhansl, 1993), spectral centroid (Grey, 1977; Krumhansl, 1989), attack and decay times (Lakatos, 2000; Krumhansl, 1989), pitch strength (McAdams et al., 1995; Lakatos, 2000), attack synchronicity (Grey, 1977; Grey and Gordon, 1978), spectral flux (Grey, 1977; Grey and Gordon, 1978; McAdams et al., 1995) and noisiness (Lakatos, 2000). In addition, Emiroglu (2007) also included further features in: spectral irregularity,

spectral spread and spectral slope (Krumhansl, 1989; McAdams et al., 1995), harmonic proportion (McAdams et al., 1995; Lakatos, 2000), overtone synchronicity and fluctuation strength (Grey, 1977; Grey and Gordon, 1978; McAdams, 1999), and roughness (Terhardt, 1974; Pressnitzer and McAdams, 2000).

The high number of features found to be of importance to timbre perception in the two meta-studies (McAdams et al., 2006; Emiroglu, 2007) leads to two assumptions. Firstly, in accordance with the main criticism of timbre studies, the results are highly dependent on the input stimuli chosen by the researcher (Caclin et al., 2005). Secondly, the number of correlating audio features with perceptual dimensions can also indicate the multidimensional nature of timbre itself. The use of MDS, and the subsequent correlation of perceptual dimensions with audio features, further highlights the unique characteristic of timbre, in that it operates in multiple dimensions.

Semantic dimensions

Similar to the discrepancies of the non-verbal method, the verbal approaches also exhibit similar problems. These studies have presented inconsistent results where little consensus is reached regarding the definitions of individual adjectives. The problem potentially originates from the use of different adjectives between studies, as well as the common problems of timbre research (i.e. the dependence of the results on the input stimuli, and the subjectivity of the listener's perception).

von Bismarck (1974) presented a semantic space of four prevalent opposing dimensions, in *full-empty, dull-sharp, color-colorless* and *compact-diffused*, while Pratt and Doak (1976) proposed a three-dimensional space that would feature the dimensions of *bright-dull, warm-cold* and *rich-pure*. Other studies have found the dimensions of *bright-gloomy, clear-dark, hard-soft, delicate-sharp, full-narrow, gloomy-clear, harsh-delicate* and *wide-narrow* to be of importance (Stepánek, 2006; Moravec and Štěpánek, 2003). From these resulting, salient dimensions, there exists a degree of variation between the descriptive terms.

Theoretically, the problem of diverging terms in the semantic dimensions of timbre can also arise due to a plethora of different instruments and sounds in a testing environment. This is a potential reason that subsequent research has focussed on timbral characteristics relating to specific instrument sounds. A number of studies have focused solely on the se-

mantics of an individual instrument, such as the violin (Fritz et al., 2012; Saitis et al., 2012), piano (Traube et al., 2008), pipe organ (Disley and Howard, 2004), saxophone (Nykänen et al., 2009) and clarinet (Barthet et al., 2010). As a result, research has also expanded into polyphonic sounds (Alluri and Toiviainen, 2010), and the acoustic qualities of concert halls (Lokki et al., 2011). This move towards descriptive terms that are associated with an instrument, a polyphonic sound or the acoustics of a room, suggests it is possible to uncover the underlying meaning of such terms in other areas. One such example is the operations a sound engineer follows to achieve certain timbral modifications through the use of audio effects.

2.2 Semantic terms

Semantic terms have been a common way of describing timbre from early studies, such as von Helmholtz (1912). Howard and Tyrrell (1997) provide a presentation of the remarks made by von Helmholtz (1912), which is broken down into a series of rules, called Helmholtz rules. Here, specific timbral adjectives are attributed to either spectral information or specific instrument types (Table 2.1). Each of the Helmholtz rules presented in Table 2.1 is associated not only with a timbral descriptor, but also with the timbre of an instrument. This suggests that the spectral structure and information can be interpreted non-exclusively for both instrument types and timbral adjectives, and justifies the approaches of timbre analysis for both timbre and semantic spaces.

While von Helmholtz (1912) balanced his interpretation of descriptors between the source (instrument) and the spectral information, Lichte (1941) deconstructed the timbre of complex tones into three independent semantic dimensions (*fullness*, *roundness* and *brightness*). However, as von Helmholtz (1912) highlighted an existing interconnection between semantic terms and instruments, there appears to be an underlying point of convergence between the two researchers. In attempting to find a correlation between non-verbal and verbal methods of timbre analysis, Zacharakis (2013) finds a strong positive correlation between the semantic and timbre spaces in his gathered datasets.

Helmholtz Rule	Frequency-domain Spectrogram	Example Timbre Descriptions	Example Instruments
1	f_0 dominates	Pure, soft, pleasant, dull at low pitch, free from roughness	Tuning fork, wide stopped organ flues
2	Harmonics dominate	Sweet, soft, rich, splendid, dark, dull, less shrill, bland	French horn, flute, tuba, open organ flues, soft sung sounds
3	Odd harmonics dominate	Hollow, nasal	Clarinet, narrow stopped organ flues
4	Striations dominate	Cutting, rough, bright, brilliant, shrill	Other reed instruments, other brass instruments, other sung sounds, bowed instruments, harmonium, organ reeds

Table 2.1: von Helmholtz (1912) rules as illustrated by Howard and Tyrrell (1997). Each rule corresponds simultaneously to spectrogram structure, timbral quality and corresponding instruments.

2.2.1 Origin of descriptive terminology in music

The origin of the terminology used to describe sound and music has been widely researched in the psychoacoustics (Wake and Asahi, 1998) and neuroscience (Koelsch, 2011) literature, as well as sound engineering (Toulson, 2006). Based on the research of Wake and Asahi (1998); Koelsch (2011) and Toulson (2006), it is possible to compound the categories into four layers of groups (Figure 2.7).

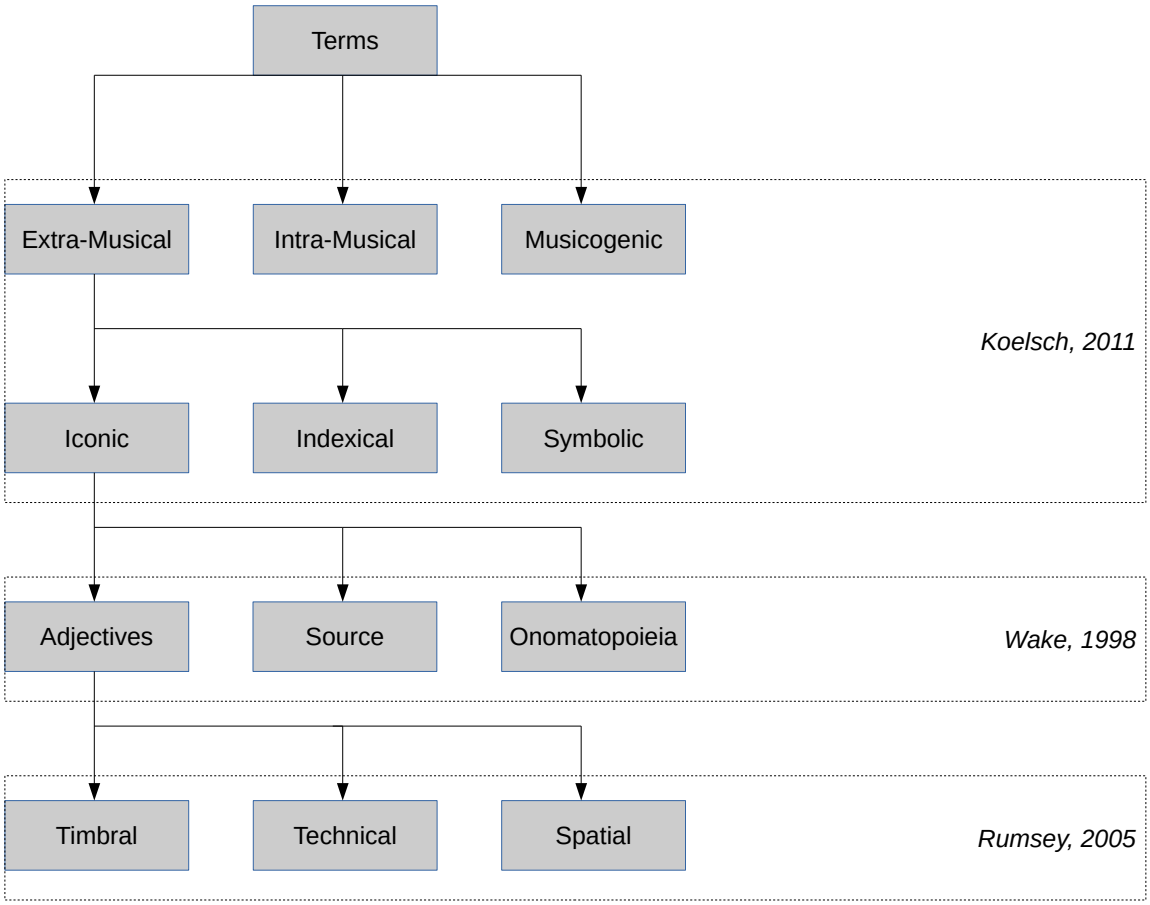


Figure 2.7: Taxonomy of descriptive terminology related to music and sound. Timbre researchers follow this taxonomy in order to choose the semantic terms used in their studies.

The first layer of Figure 2.7 is developed by Koelsch (2011). In considering musical semantics a part of language processing, the terminology can be broken down into three categories. Firstly, extra-musical meaning, which is associated with terms that are not directly connected to the structure, melody, progression or rhythm of the music. The second category is intra-musical meaning, which arises from the interpretation of references associated with the music itself, such as the rhythmic patterns and harmonic sequences. Finally, the third category is that of musicogenic meaning, in which the subject derives a personalized response, physical, emotional or self-related, to a sound. This category is predominantly theoretical, since acquisition of data and interpretation differs greatly between individuals and is very difficult to quantify.

Extra-musical meaning

The second layer of Figure 2.7 stems from the extra-musical category, which can be broken down into three further groups. Firstly, iconic syntax, in which musical patterns that resemble sounds, quality of objects or abstract concepts are used. Secondly, indexical syntax, which relates to sound patterns that index the presence of a psychological state, such as an emotion or an intention. Finally, symbolic syntax, which emphasizes the symbolic qualities of music. An example of symbolic meaning is the case of an individual listening to the national anthem of their country.

Iconic extra-musical meaning

The third layer of Figure 2.7 originates from the iconic sub-category of extra-musical meaning. This layer of categories was introduced by Wake and Asahi (1998) and can be considered a continuation of Koelsch (2011). The first category of this layer is onomatopoeia (roughly translated to "name crafting"), and relates to the creation of new sounds or words in order to describe the input stimuli. The second category, source, relates to the identification of the source of the perceived sound. This can be the type of instrument and its qualities (for example "acoustic", "metallic" or "electric") or the place of origin of environmental sounds and noises. Finally, the third category, adjectives, describes the use of adjectives to better describe the perceived sound and includes words such as *bright*, *warm* and *clear*.

It should be stated that the categorisation of Wake and Asahi (1998) presents classes that are not exclusive from each other. The task of recognizing an instrument, describing its sound through the creation of words/sounds, and using adjectives to define its timbre, can all be performed for the same sound. The three processes therefore can be performed simultaneously and are independent of each other. This also correlates with von Helmholtz (1912) and Table 2.1, where instruments (source) and semantic terms (adjectives) share similar characteristics.

2.2.2 Semantic terms in sound engineering

The disciplines of sound engineering and music production also entail the use of semantic terms to describe the timbre of a sound. Huber and Runstein (2013) explain that micro-

phone selection can influence the timbre of the sound being recorded. This is not limited to the type of microphone that is used, but it is also its placement that can alter the perceptual characteristic of the sound. This point is reinforced by the suggestion of Owsinski (2009) regarding recording techniques, where the author makes use of descriptive terms to compare two microphones ("*The KM56 has a brighter, more aggressive quality than the KM54*"), but also explains how positioning, and more specifically the proximity effect (when a microphone is placed closer to a source) is able to add *warmth* and *fullness* to the sound.

This method allows the sound engineer/music producer to explain a technical and complicated process through the use of simple terms, which are understood by other people in the field. This can be very helpful in the context of mixing and mastering music, where target sounds can be explained without the need of detailing every single parameter used to achieve the resulting timbral characteristic (Rumsey, 2005).

Figure 2.7 includes an additional layer, which considers a further categorisation of adjectives to Wake and Asahi (1998). The three further categories include adjectives that relate to: technical quality, spatial quality and timbral quality. This sub-category relates to the operational actions of sound engineers, suggesting that despite the disagreements on definitions of specific terms, there are instances where the meaning of some words is more agreed upon. Rumsey (2005) argues that a degree of agreement exists in the technical and spatial categories. Words such as *compressed* and *noisy*, even though they do not define a specific extent (how *compressed* or *noisy* a signal is), are not only agreed upon, but are even detectable through analysis of the signal. Similarly, terms regarding the spatial quality of the audio can be traced to certain aspects of production, such as the use of reverb or panning used to achieve adjectives like *deep, wide* and *up-front*. However, an agreement on timbral adjectives is elusive, since such terms relate to imagery, feelings and visualizations, similar to the extra-musical category of terms proposed by Koelsch (2011).

Izhaki (2013) further suggests that the origin of the problem exists due to the use of different languages between individuals dealing with sound. For instance, sound engineers will discuss their processing in terms of frequency bands, measured in Hz (cut below 80Hz, or boost 4kHz) while musicians will talk in terms of notes (low E note). Furthermore, the use of certain adjectives for a prolonged period of time has resulted in multiple possible interpretations and processing strategies to achieve such a result. For example, different

equalisation settings are able to achieve the same timbral adjective (Katz, 2003). Similarly, different audio effects can produce the same timbral characteristic, for instance both harmonic distortion and subtle compression can lead to a *warm* sound (Hood, 1997).

Despite the disagreement on terminology, engineers and producers have been using descriptive terms in order to define the timbral characteristic of specific sounds and processes. These can be found in most of the existing literature regarding recording, mixing and mastering audio, where parameters of audio effects are associated with semantic terminology (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005).

Toulson (2006) has discussed the requirement for a universal audio terminology, which will bypass the need for technical parameters in the context of audio effects. He suggests novel parameters can be created, based on the concept of semantic terms. However, due to the lack of progression and correlation in timbre research, Cartwright (2016) suggests that the establishment of a universal dictionary remains largely unresolved.

Models using semantic terminology

Synthesis methods The inherent use of semantic terms in music production has led many researchers to attempt to create systems and models that are able to transform an input sound without the need for technical parameters. Brookes and Williams (2007); Williams and Brookes (2009); Brookes and Williams (2010) produced research using a technique known as audio morphing (Slaney et al., 1996) for a system that could morph an input sound using three terms (*brightness, warmth, softness*) independently of each other.

Using perceptually-motivated morphing the authors were able to transform timbral characteristics through the manipulation of audio features. For *brightness*, Brookes and Williams (2007) manipulated the spectral centroid, an audio feature showing correlation with the perception of the term (Beauchamp, 1982). For *softness*, existing research (Disley and Howard, 2004; Freed, 1990; Lakatos, 2000; Fletcher, 1999; Stepánek, 2006) reveals a correlation of the term with the inharmonicity and attack time. However after preliminary testing Williams and Brookes (2009) found the effect of the inharmonicity to be negligible and only the attack time was used. Finally, for the manipulation of *warmth* (Brookes and Williams, 2010), the amplitude of the three first harmonics was considered responsible for the timbral characteristic (Ethington and Punch, 1994). Despite the results of the audio

morphing papers (Brookes and Williams, 2007; Williams and Brookes, 2009; Brookes and Williams, 2010), and their use of correlating features that influence specific timbral qualities, Zacharakis and Reiss (2011) criticised the approach regarding its independent modification of each term. This finding also arises from the literature, since by manipulating the first three harmonics in Brookes and Williams (2010), the fundamental frequency was affected, which according to Marozeau et al. (2003); Marozeau and de Cheveigné (2007) contributes to the perception of *brightness*. Therefore, the processing cannot be considered independent.

Application of semantic terms in a synthesis context has also proved popular in recent studies. Miranda et al. (2000) brought up the issue that using the conventional taxonomy of musical instruments is limiting for synthesizer usage, and proposed an alternative taxonomy that took into account the textural and timbral characteristics of a sound. To this end, Ethington and Punch (1994) presented SeaWave, a synthesizer in which the usual technical parameters of envelope manipulation, filter controls, resonance, modulation and different types of waveforms were replaced by semantically meaningful terms. Similarly, Howard et al. (2007) performed listening tests in order to find the most antithetical semantic terms, which would then be used to replace the interface's technical parameters for a synthesizer.

Professional equipment Professional audio equipment is dominated by technical parameters, as is presented in Figure 2.8 with the Master Hills Mastering Compressor. From the figure it can be deduced that the equipment requires a firm background in sound engineering in order to be appropriately used. Parameters such as 'gain' might not be foreign to novice users, but terms such as 'attack', 'recover', 'ratio' and 'threshold' might be unknown to them.

Figure 2.8: The Shadow Hills Mastering Compressor.

Meanwhile, for a number of years there has also been a trend in the industry of making use of semantic terms in order to enhance interactivity with the equipment. An example of this hybrid approach is the Empirical Labs Distressor compressor (Figure 2.9), which includes a 'nuke' option on the ratio setting that corresponds to a highly compressed result.

Figure 2.9: The Empirical Labs Distressor.

Manufacturers of software-based audio effects, such as Waves, have taken this approach to completely replace technical parameters and instead provide the user with a single dial that controls a specific timbral quality. This is known as the OneKnob series (Figure 2.10),

and includes effects that are able to make a sound *brighter*, *phatter* or *wetter* among more options. In this instance the user does not even need to be aware as to the type of audio effect that is being implemented to achieve the timbral characteristic and can simply adjust the amount of transformation that is deemed desirable.

Figure 2.10: The Waves OneKnob series of audio effects.

The role of equalisation

Despite disagreement on descriptive terms, there is one area regarding the use of timbral adjectives where the literature for sound engineering, especially in the context of mixing and mastering, appears to be converging. One of the first instances of the relationship between equalisation and timbral adjectives was presented by Kulka (1972), who split the frequency range into a number of bands and described the effect that changing the gain of each band will have on the resulting sound (Table 2.2).

Frequency (Hz)	Timbral Characteristic
31	Rumble, "chest"
63	Bottom
125	Boom, thump, warmth
250	Fullness or mud
500	Honk
1000	Whack
2000	Crunch
4000	Edge
8000	Sibilance, definition, "ouch"
16000	Air

Table 2.2: Timbral adjectives with their respective frequency ranges according to Kulka (1972). This table is one of the earliest examples of how semantic terminology is understood by sound engineers, and it provides guidance as to how these terms can be manipulating through the use of equalisation.

In addition to providing the results of changes in the frequency spectrum, in the form of timbral characteristics, Kulka (1972) also provided a set of rules for the correct handling of equalisation, presented in Table 2.3. What is interesting about this table is that two of the rules (1, 2) are associated with a specific timbral adjective (*muddy* and *honky*). In addition, Kulka (1972) suggests that whilst cutting frequencies will result in a "better" sonic result, boosting frequencies will result in the sound appearing "different". This can mean that equalisation is responsible for the timbral shift between the original sound and the processed output.

Kulka Rule	Comment
1	If it sounds muddy, cut some at 250 Hz
2	If it sounds honky, cut some at 500 Hz
3	Cut if you are trying to make things sound better
4	Boost if you are trying to make things sound different
5	You can't boost something that's not there in the first place

Table 2.3: Equalisation rules according to Kulka (1972). These rules display the effectiveness of conveying guidance by using semantic terminology.

The relationship between equalisation and timbral adjectives is found in recent audio mixing literature, with Owsinski (2013) suggesting that the main tool of the mixing engineer for achieving certain timbral adjectives (to make a sound *brighter, fatter, bigger* or *clearer*) is equalisation.

Similarly, Izhaki (2013) agrees on the assumption that semantic terms are associated with the frequency range, but suggests that the characteristics of the input audio need to be taken into account. Finally, Katz (2003) affirms the relationship between the descriptive terms and the frequency domain in the mastering process, presenting an illustration of the frequency range alongside timbral adjectives associated with the different ranges (Figure 2.11).

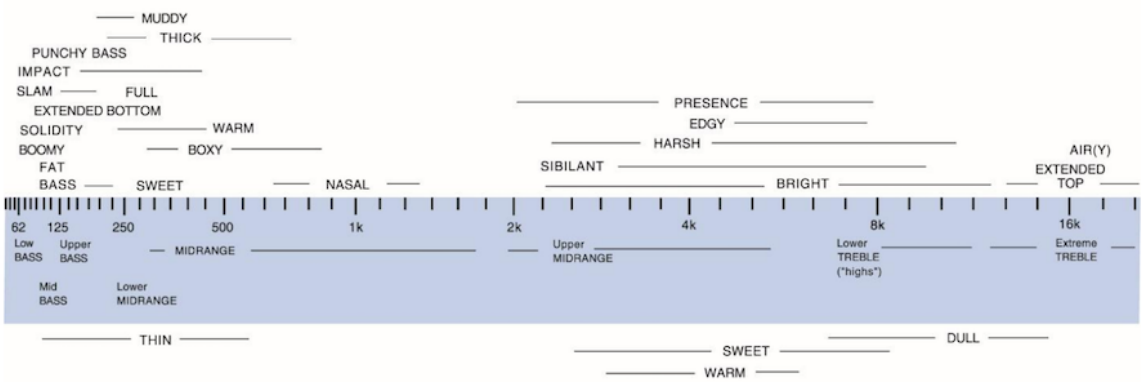


Figure 2.11: Katz (2003) proposed distribution of semantic terms over the frequency domain. The placement of the terms above or below the frequency axis correspond to whether the term can be achieved by cutting or boosting that range. The illustration also provides a translation between semantic terminology and different parts of the frequency spectrum (BASS, MIDRANGE, TREBLE).

This practical view of sound engineers, reinforced by existing literature in mixing and mastering techniques, suggests that a majority of terms is achievable through equalisation. De Man (2017) produced a comparative table highlighting the frequency ranges considered by professional sound engineers to be responsible for achieving certain timbral adjectives (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005), a sample of which is presented in Table 2.4. This further suggests that semantic terms are widely associated with frequency components, and by extension equalisation, and it can therefore be considered that the

specific processor is essential for timbral transformations.

Term		Range	Reference
air		5–8 kHz	(Gibson, 2005, p. 119)
		10–20 kHz	(Coryat, 2008, p. 99)
		10–20 kHz	(Izhaki, 2013, p. 211)
		11–22.5 kHz	(Owsinski, 2013, p. 26)
		12–15 kHz	(White, 1999, p. 103)
		12–16 kHz	(Katz, 2003, p. 43)
		12–20 kHz	(Coryat, 2008, p. 25)
		12–20 kHz	(Cousins and Hepworth-Sawyer, 2013, p. 108)
		12–20 kHz	(Waddell, 2013, p. 86)
bright		2–12 kHz	(Katz, 2003, p. 43)
		2–20 kHz	(Izhaki, 2013, p. 211)
		5–8 kHz	(Owsinski, 2013, p. 119)
distant	lack of	200–800 Hz	(Gibson, 2005, p. 119)
	lack of	700–20 000 Hz	(Izhaki, 2013, p. 211)
	lack of	4–6 kHz	(Owsinski, 2013, p. 25)
	lack of	5 kHz	(Huber and Runstein, 2013, p. 484)
presence		800–12k Hz	(Gibson, 2005, p. 119)
		1.5–6 kHz	(Coryat, 2008, p. 24)
		2–8 kHz	(Katz, 2003, p. 43)
		2–11 kHz	(Izhaki, 2013, p. 211)
		2.5–5 kHz	(Huber and Runstein, 2013, p. 484)
		4–6 kHz	(Owsinski, 2013, p. 25)
warm, warmth		90–175 Hz	(Gibson, 2005, p. 26)
		100–600 Hz	(Izhaki, 2013, p. 211)
		200 Hz	(Huber and Runstein, 2013, p. 484)
		200–800 Hz	(Owsinski, 2013, p. 119)
		200–500 Hz	(Cousins and Hepworth-Sawyer, 2013, p. 105)
		250–600 Hz	(Katz, 2003, p. 43)

Term	Range	Reference

Table 2.4: Sample of timbral adjectives in the sound engineering literature and their corresponding frequency ranges, as gathered and presented by De Man (2017).

Crowd-sourced systems for research on timbral adjectives

In order to comprehend the process sound engineers undertake in their timbral manipulation, a number of researchers have turned towards crowd-sourcing practices. The Interactive Audio Lab at Northwestern University has proposed a series of different audio effects (Zheng et al., 2016), including equalisation (Cartwright and Pardo, 2013), reverb (Seetharaman and Pardo, 2014a) and compression (Zheng et al., 2016), capable of extracting the users' definition of semantic terms. All the systems were hosted in an web browser and subjects were asked to visit a URL in order to participate in the experiment. The aim of the systems is to uncover the underlying attributes of timbral adjectives, the confidence of users in the terms' definition, the degree of similarity between terms, as well as establishing a language basis that will minimize or eliminate the need for technical parameters in audio effect parametrisation.

Alternatively, Stables et al. (2014) proposed a system that could extract information from a Digital Audio Workstation (DAW) in order to comprehend the processing that takes place when semantic transformation is performed. For this reason, four audio effects (equalisation, compression, distortion and reverb) were built in standard format, to operate within any given setup. These audio effects are used for collecting data, which includes the audio features before and after the audio effect has processed the sound. This method of data gathering allows for a more accurate representation of audio engineering concepts and practices, since it captures information from engineers within their familiar environment, manipulating audio effects the same way they would when mixing or mastering. For the extraction of information, the Semantic Audio Feature Extraction (SAFE) project implemented the LibXtract feature extraction framework developed by Bullock (2007), which is capable of extracting 80 features. The library includes features that are relevant to the dimensions of timbre, in the log attack time, spectral centroid and spectral flux among more. A comparative review performed to assess the attributes of audio feature extraction libraries (Moffat et al., 2015) has found that LibXtract provides extensive real time

extraction methods, as well as wide coverage of audio features.

In addition to feature information, which is extracted before and after the audio effect has been applied to the signal, the SAFE plugins also retrieve the parameter settings at the time of the extraction process, as well as additional metadata. These include context relevant information such as the genre and musical instrument being processed, in addition to metadata regarding the user processing the audio, such as their location, age and experience in the field. Figure 2.12 details the operational process of the SAFE audio effects.

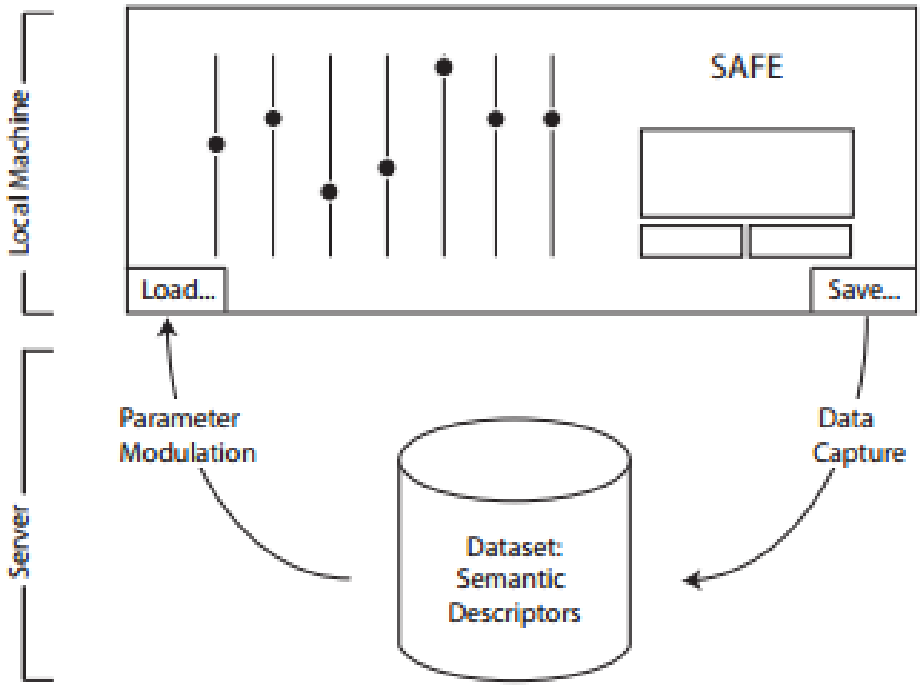


Figure 2.12: The SAFE project structural architecture. Here the user works on a local machine, and the mechanism allows them to save parameters by providing a semantic descriptor, which are then stored in a server. The users also have the option of loading parameters by typing a semantic term.

2.3 Conclusion

Timbre research has gone to great depths in order to measure perceptual differences between sounds. The non-verbal, MDS approach has been able to uncover that perceptual

dimensions are correlated with three types of audio features, namely the temporal, spectral and spectrotemporal envelopes of the signal. Similarly, the development of statistical methods regarding these dimensions has allowed for a deeper evaluation of timbre. However, the main criticism of using the official definition of timbre (ANSI, 1973) has not been bypassed, and still poses a problem to the research. Additionally, the perceptual nature of timbre, lack of sufficiently correlated audio features, and the dependence of the test results on the choice of input stimuli, adds an additional layer of complexity to the task.

On the other hand, the verbal approach, including VAME and semantic differential techniques, with the use of dimensionality reduction, share some of the same problems with the MDS approach. Mainly, the results of these studies are still dependent on the input stimuli chosen for the testing process. Moreover, an additional layer of complexity is added to the process due to the requirement of the researchers to choose the appropriate timbral adjectives for testing purposes.

Another problem that has not been addressed is the effect that timbre might have on the other auditory dimensions, as was initially suggested by Melara and Marks (1990). Research has shown that even in cases where timbre does not objectively affect the pitch or loudness of a signal, it can still alter the perception of these dimensions, as was displayed in Vurma et al. (2011), where pitch perception was affected by changes in timbre. This particular finding can cause the original definition of timbre to unravel, since one of the dimensions that should be theoretically unaffected is altered, a point initially made by Krumhansl (1989). Also in contrast with the ANSI (1973) definition of timbre, the fundamental frequency, for instance, was found to correlate with the perception of *brightness* (Marozeau et al., 2003; Marozeau and de Cheveigné, 2007), suggesting that a *higher* pitch could potentially be described as *bright* when compared to a lower pitch. Similarly, studies (Kuwano and Namba, 2001) have found a correlation between the total loudness and descriptors such as *powerful*, another dimension that is excluded from timbre research, according to ANSI (1973).

While this might make the research of timbre more challenging, it also opens up its effect to further interpretations. An aspect arising from timbral characteristics, the use of semantic terms, can potentially provide meaningful information regarding timbral transformation, within a set context, such as sound engineering and the areas of recording, mixing and mastering.

Sound engineers make use of terms in order to characterise the timbre of a sound, which stem from the operational actions necessary to perform a timbral transformation. These transformations are performed through the use of audio effects. As is depicted in Figure 2.13, processing modules (audio effects) can be placed on the input audio in order to alter its timbral characteristics at the output stage.

Figure 2.13: Use of processing modules to alter the timbre of an input sound. Engineers will place different audio effects in series in order to construct processing chains, which allow more complex transformations over the timbre of the input audio.

Furthermore, the existing literature suggests that equalisation is the most commonly used method for altering the timbral characteristic of a sound (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005). Through applying gain on certain frequency ranges, the input sound can take on different timbral representations, as is depicted in Figure 2.11. For instance, boosting the frequency components in the range of 2-8 kHz will result in adding *presence* to a sound, or a similar boost in the range of 4-8 kHz would make the sound more *edgy*.

The study of audio effects altering timbre is closely related to the semantic space approach in timbre analysis. However, instead of grading audio samples by the semantic terms used to describe them, the changes in audio effect parametrisation can produce a method of grading on its own. While studies on timbre research focus on unchanging audio features, aiming at the correlation between audio features and the uncovered perceptual dimensions, the use of audio effects will alter the audio features from the input to the output stage. That is a point made by Enderby and Stables (2017), where *differences* in audio features between input and output stages can also be responsible for deriving specific semantic terminology.

Chapter 3

Processing chain analysis

3.1 Introduction

Audio effects are an integral part of sound engineering, allowing engineers and producers to alter the perceptual characteristics of an input sound. However, in order to easily and efficiently describe the complex process of optimising the parameters of each audio effect, producers and engineers will make use of timbral adjectives. Based on this premise, a number of different projects have conducted research focusing on the use of audio effects in achieving these descriptive terms, such as the SAFE (Stables et al., 2014) and SocialFX (Zheng et al., 2016) projects.

The two aforementioned projects investigated the application of audio effects for creative purposes, i.e. for altering the timbral quality of an input sound. In order to explore the impact of individual audio effects on the perceptual characteristics of an input sound, research on individual audio effects such as equalisation (Stasis et al., 2015; Stasis, Stables and Hockman, 2016; Cartwright and Pardo, 2013), distortion (Enderby and Stables, 2017), compression (Maddams et al., 2012) and reverberation (Seetharaman and Pardo, 2014a,b) has been performed and has produced a further insight on the use of these applications.

However, sound engineers will usually not implement creative changes through a single audio effect, but will commonly place audio effects in series to create what is known as a processing chain (Figure 3.1). These chains of audio effects allow for a more complex and versatile method of transforming the input sound. While cascading these processing modules in series might be a simple task for sound engineers and music producers, the

analysis of their impact on the input signal presents a non-trivial problem. The complexity of the individual audio effects, which can be considered multidimensional processing units, and the combination of linear and non-linear systems, which are non-commutable, suggest a very large number of possible combinations.

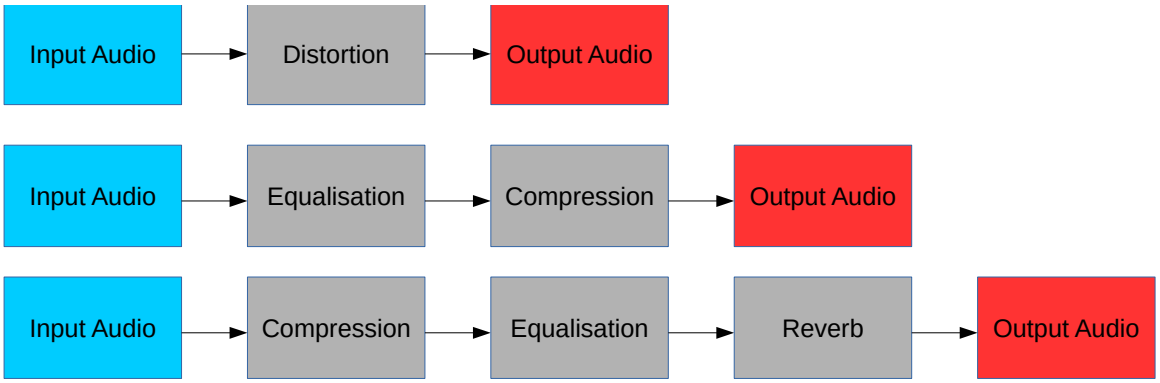


Figure 3.1: Examples of processing chains. Here, the blue boxes denote the input audio, the red boxes the denote transformed, output audio and the grey boxes depict the processing modules used to perform the transformation.

Still, research on processing chains has been performed, focusing mainly on the ordering of the effects in a chain. Considering the two main effects in the sound engineer's disposal, equalisation and compression, various suggestions have been made as to the relevant placement between the two. Pestana (2013) for example, argues that compression placed before equalisation will allow for more control over the tonal balance of the track, while the alternate placement ensures better control over the dynamic range. This discussion has also led to novel systems being developed, such as the processing unit proposed by Wise (2009), which combines the manipulation of both dynamics and frequency components within a single framework.

Despite the in-depth investigation of both the semantic attributes of individual audio effects and the characteristics of processing chains, there has been a lack of research that focuses on the use of processing chains to alter timbral characteristics. By designing an open ended experiment it is possible to derive answers as to the use of processing chains for achieving timbral adjectives. For example, are audio effects placed in series preferred over single audio effects for performing timbral transformations? What are the most common processing chains used to alter timbre? What terms correlate with which audio effects?

What number of audio effects (length of chain) is required to perform these tasks? In addition, the study of processing chains can be expanded by providing a further layer of similarity between chains. In that context, chains that are used to derive the same group of semantic terms can be considered more similar to each other.

Finally, through this analysis it is possible to deconstruct the process sound engineers follow in placing audio effects in series to achieve individual timbral transformations. Based on this knowledge a processing chain recommendation system can be produced that will provide more creative options for sound engineers in altering the timbre of an input sound, as well as aiding novice users in understanding what audio effects are necessary for performing these creative tasks.

3.2 Experiment design

In order to evaluate the use of processing chains in the context of timbral transformations, an open ended experiment is designed. Subjects are provided with audio samples and timbral adjectives, and are asked to perform the given timbral transformations using a range of audio effects. The tests were performed through a web interface, in which subjects were provided with a URL and asked to participate from their home studios. While implementing this method of testing introduced additional variables, caused by the listening environment, a larger collection of data can be obtained. This practice is common in audio research (Cartwright and Pardo, 2013; Birnbaum, 2004) given that the participants have undergone a screening process (Cartwright et al., 2016). Subjects were chosen from an undergraduate sound engineering course, having studied mixing concepts and audio effects usage for at least two years. The subjects were asked to follow a set of instructions for a predetermined period of time. To tailor the testing process to their availability, subjects could choose a testing duration, ranging from 5 to 30 minutes.

Subjects were supplied with the SAFE plug-in suite, comprising four audio effects: (1) parametric equalisation, (2) dynamic range compression, (3) non-linear distortion and (4) algorithmic reverberation. The initial format of the SAFE audio effects, which allowed them to operate within any DAW, was altered to allow easier crowd-sourcing capabilities through the JSAP audio plugin framework (Jillings, Wang, Reiss and Stables, 2016), moving the audio effects to a web browser. The implementation of the audio effects still uses

the architecture of the SAFE project, where features are extracted before and after processing, along with the parameter values of the audio effect and relevant metadata. Whenever a processing chain is submitted, the server can extract audio features between each node in the chain, using the JS-Xtract feature extraction library (Jillings, Bullock and Stables, 2016). Furthermore, in order to not tamper with the creative nature of the task, no restrictions were placed on the order of the audio effects, the length of the processing chain and the number of instances an effect could appear in the chain.

Alongside the audio effects, the subjects were given a number of predefined audio samples. These audio samples were selected from the Mixing Secrets Library via the Open Multitrack Testbed (De Man et al., 2016). In order to examine a wide range of audio samples, various instruments across different genres were chosen for the testing phase. The instruments included *acoustic guitar*, *bass guitar*, *drums* (mixed), *electric guitar*, *piano*, *saxophone*, *violin* and *vocals*, as well as *complete mixes*, which were mixdown samples of the individual tracks. These samples covered five genres, namely *Reggae*, *Folk*, *Rock*, *Hip-Hop* and *Jazz*.

In turn, a set of descriptive terms that the subjects would be asked to achieve needed to be chosen. For this purpose, the descriptors were selected from the SAFE project dataset (Stables et al., 2016). The two attributes that were taken into account for their selection was the frequency by which they appeared in the SAFE dataset, and their generality across audio effects. In this context generality is defined as the extent to which the term can be achieved through different audio effects. This was to ensure that both terms that were associated with a specific audio effect and terms that could be achieved through multiple audio effects were used (Stables et al., 2016). The resulting adjectives included: *air*, *boom*, *bright*, *close*, *cream*, *crisp*, *crunch*, *damp*, *deep*, *dream*, *fuzz*, *punch*, *room*, *sharp*, *smooth*, *thick*, *thin* and *warm*. Due to an overlap between the most frequently used terms and terms exhibiting the highest generality, the total number of terms for this experiment dropped to 18.

In total 47 subjects participated in the tests, and 178 submissions were provided. The subjects were aged between 18-50, had various levels of experience (0-32 years) in professional audio production, and none displayed any hearing impairment. In each session, the subjects were provided with the audio samples and timbral changes that they needed to perform. The genre, instrument and descriptive term were randomly chosen for each

session.

From this dataset it is possible to perform a number of measurements for understanding both the role of individual audio effects in processing chains, presented in Section 3.3, as well as the similarity between processing chains, presented in Section 3.4. In addition Section 3.5 describes the manner in which a processing chain recommendation system based on semantic cues can be implemented.

3.3 Single effect analysis

Of the four available plugins, 124 equalisation, 72 compression, 57 reverb and 40 distortion instances were used. Around half of the submissions comprise single effect chains. The longest processing chain present in the dataset features four plugins, a length capable of providing 256 possible permutations. Tables 3.1, 3.2 and 3.3 display the number of descriptors, instruments and genres that the subjects were asked to provide processing chains for.

3.3.1 Effect generality

The generality of an effect is defined as the extent to which it can be used across different contexts (instruments, genres or descriptors). A low generality score is associated with a plugin appearing in a single context instance, for example if the distortion effect is applied only to a snare sample. On the other hand, if the effect is applied to all the instances, for example compression being applied to all the different music genre samples, the generality score will be high. Equations 3.1 and 3.2 are used to calculate a generality score from the data for a given contextual term. These were adapted from Stables et al. (2016).

$$g_i\left(p\right) = \frac{2}{D-1} \sum_{d=0}^{D-1} d\text{sort}\left(x\left(d,p\right)_i\right) \tag{3.1}$$

$$x\left(d,p\right)_i = \frac{n_p\left(d,i\right)}{\sum_{d=0}^{D-1} n_p\left(d,i\right)} \tag{3.2}$$

In Equations 3.1 and 3.2, $g_i(p)$ is the generality of a plugin p on instrument i and D is the total number of descriptors. $n_p(d,i)$ is the number of plugin occurrences on descriptor

d and instrument i. As was performed in Stables et al. (2016) a sorting ($sort$) algorithm is used to find the weighted mean of the sorted distribution. The measurements refer to the range of plugins that are used to process a given instrument. The generality g will be low if a plugin occurs for a small number of instruments and increases monotonically with the number of instruments.

The measurements show that the distortion is the least general effect, whilst the EQ displays a consistently high generality score. On the other hand, the compressor appears to be more general for the *piano*, *drums* and *complete mixes* samples (Table 3.2), suggesting that it is able to surpass the timbral adjective and genre, and be implemented irrespective of them. From Tables 3.2 and 3.3, it is clear that both instrument and genre have little impact on the choice of audio effects, with all plugins displaying similar generality scores across the various types of instruments and genres.

Table 3.4 shows the cumulative generality scores for an effect being used across the different contexts. A low score here indicates that there is a high likelihood of the effect being used. The results further suggest that the genre and instrument play a relatively small role in the selection of a plugin within a processing chain. The only attribute that exhibits an impact is the descriptor type for the use of distortion and reverberation. While equalisation and compression appear to be more evenly spread out, achieving a higher generality score, the results for reverberation and distortion are more specific, and therefore the two effects are used sparingly.

3.3.2 Effect salience

As the web-based version of SAFE using the JS-Xtract library is capable of extracting a number of audio features between the various processing phases of the chain, the impact of each plugin can be measured. This is performed by calculating the Euclidean distance between audio features before and after processing in the chain (Equation 3.3).

$$distance(\mathbf{fb}, \mathbf{fa}) = \sqrt{\sum_{i=1}^{N} (fa_i - fb_i)^2} \tag{3.3}$$

where fb are the features before reaching a plugin in the chain, and fa are the features after the effect has processed the audio sample. Each feature is vector normalised against

Desc.	#	EQ	Comp.	Dist.	Revrb.
air	7	4 (0.60)	3 (0.40)	0 (-)	3 (0.40)
boom	7	7 (0.86)	4 (0.30)	0 (-)	0 (-)
bright	10	10 (0.73)	3 (0.33)	1 (0.00)	3 (0.33)
close	13	11 (0.64)	5 (0.40)	1 (0.00)	5 (0.67)
cream	9	8 (0.81)	4 (0.38)	0 (-)	3 (0.17)
crisp	9	9 (0.58)	3 (0.44)	0 (-)	1 (0.00)
crunch	15	5 (0.34)	7 (0.61)	14 (0.78)	2 (0.14)
damp	8	4 (0.75)	1 (0.00)	1 (0.00)	9 (0.83)
deep	9	9 (0.82)	3 (0.11)	0 (-)	2 (0.17)
dream	9	4 (0.50)	1 (0.00)	1 (0.00)	9 (0.82)
fuzz	11	2 (0.25)	0 (-)	11 (0.64)	1 (0.00)
punch	9	7 (0.86)	9 (0.71)	1 (0.00)	0 (-)
room	13	4 (0.50)	2 (0.17)	1 (0.00)	13 (0.72)
sharp	7	7 (0.86)	5 (0.53)	1 (0.00)	0 (-)
smooth	9	6 (0.67)	5 (0.48)	2 (0.20)	3 (0.40)
thick	9	8 (0.56)	5 (0.20)	3 (0.50)	1 (0.00)
thin	11	11 (0.73)	1 (0.00)	1 (0.00)	2 (0.17)
warm	13	11 (0.82)	5 (0.40)	2 (0.17)	3 (0.11)

Table 3.1: Audio effect generality across descriptors. # corresponds to the total number of entries for each descriptor with the number of plugins applied and generality g_d across all descriptors in braces. Here, high scores suggest that the effect is specific to the corresponding plugin.

Inst.	#	EQ	Comp.	Dist.	Revrb.
Ac. Gtr	8	4 (0.43)	3 (0.29)	1 (0.00)	4 (0.43)
El. Gtr.	14	9 (0.73)	4 (0.14)	5 (0.36)	5 (0.22)
Saxo.	4	4 (1.00)	2 (0.33)	2 (0.33)	1 (0.00)
Mix	36	22 (0.68)	17 (0.40)	6 (0.15)	15 (0.48)
Bs. Gtr.	31	22 (0.65)	12 (0.36)	7 (0.08)	8 (0.27)
Vocals	21	15 (0.81)	6 (0.16)	3 (0.04)	5 (0.27)
Drums	30	21 (0.57)	13 (0.53)	9 (0.19)	8 (0.27)
Violin	17	12 (0.70)	7 (0.26)	4 (0.27)	8 (0.36)
Piano	17	15 (0.81)	6 (0.53)	3 (0.07)	3 (0.20)

Table 3.2: Audio effect generality across instruments. # corresponds to the total number of entries for each instrument with the number of plugins applied and generality g_i across all instruments in braces. Here, high scores suggest that the effect is specific to the corresponding plugin.

Genre	#	EQ	Comp.	Dist.	Reverb.
Reggae	32	22 (0.61)	13 (0.43)	7 (0.21)	9 (0.33)
Jazz	33	24 (0.66)	20 (0.65)	9 (0.10)	13 (0.49)
Hip Hop	38	25 (0.64)	10 (0.39)	12 (0.21)	12 (0.34)
Folk	22	14 (0.684)	10 (0.31)	2 (0.07)	10 (0.31)
Rock	53	39 (0.60)	19 (0.47)	10 (0.20)	13 (0.20)

Table 3.3: Audio effect generality across genres. # corresponds to the total number of entries for each genre with the number of plugins applied and generality g_g across all genres in braces. Here, high scores suggest that the effect is specific to the corresponding plugin.

Term	EQ	Comp.	Dist.	Reverb.
Genre	0.786	0.799	0.713	0.904
Instrument	0.692	0.625	0.656	0.640
Descriptor	0.766	0.631	0.291	0.464

Table 3.4: Generality of plugins against the type of term (genre, instrument, descriptor). Higher scores suggest a more general presence for the audio effect, while lower scores suggest the audio effect appears in specific instances.

all other instances of the same feature in the processing chain in order to capture the relative salience of each effect in the chain.

Table 3.5 shows the distance between features as a function of the position in the chain across all entries in the dataset. Two significant results can be deduced through this measurement. Firstly, the first processing module of the chain tends to have the most impact on the overall result, irrespective of the effect type. Secondly, non-linear effects display a higher impact on the audio features, regardless of their position in the processing chain.

Audio Effect	*1st Position*	*2nd Position*	*3rd Position*	*4th Position*
EQ	2.017	1.386	0.75	1.778
Distortion	2.654	2.068	1.592	0
Reverb	2.234	2.121	2.093	1.087
Compressor	2.19	2.383	2.449	2.366

Table 3.5: Euclidean distances of audio features according to plugin type and position in the chain. Higher values suggest a more substantial shift in audio features of the input audio.

3.3.3 Order of audio effects

Each processing chain can be considered a multi-dimensional vector, where each dimension represents a plugin instance. A range of positions $0, ..., k-1$ is defined, where each effect in the chain assumes one position in the chain. For example, in the third processing chain in Figure 3.1, the EQ assumes the first position in the chain, the compressor the second and the reverb the third.

By considering each instance in the vector as a finite state, the likelihood of a plugin appearing at position k, given the state at positions $0, ..., k-1$ can be represented using a Markov chain (Markov, 1971; Tauchen, 1986). The state vector v comprises equalisation (E), compression (C), distortion (D) and reverberation (D). Alongside the four effects that represent the states of the Markov chain, a fifth state is defined as an *empty* plugin state. The chain must start at a non-*empty* state and is considered a complete set when it reaches an *empty* state.

$$v = [\text{'empty'}, \text{'E'}, \text{'C'}, \text{'D'}, \text{'R'}] \tag{3.4}$$

$$P = \begin{bmatrix} P_{0,0} & P_{0,1} & P_{0,2} & P_{0,3} & & P_{0,k-1} \\ P_{1,0} & P_{1,1} & P_{1,2} & P_{1,3} & & P_{1,k-1} \\ P_{2,0} & P_{2,1} & P_{2,2} & P_{2,3} & \dots & P_{2,k-1} \\ P_{3,0} & P_{3,1} & P_{3,2} & P_{3,3} & & P_{3,k-1} \\ P_{4,0} & P_{4,1} & P_{4,2} & P_{4,3} & & P_{4,k-1} \end{bmatrix} \tag{3.5}$$

$$\Pr\left(A_n = p_i | A_{n-1} = p_j\right) = P_{i,i-1} = P_{i,j} \tag{3.6}$$

$$\Pr\left(A_n = p_i | A_{n-1} = p_j, ..., A_0 = p_0\right) = \prod_{n=1}^{N} P\left(A_n, A_{n-1}\right) \tag{3.7}$$

where $A_1, A_2, ..., A_n$ is a sequence of random variables with the Markov property, by which the probability of the next state depends only on the current state and not on any previous states. In addition, p_i and p_j are the elements of the state space, which are passed as input. From those the probability of state $P_{i,j}$ is derived.

Equations 3.4 and 3.5 provide the transition matrix of the chains, measuring the proba-

bility of the next plugin type given the previous plugin in the chain, as defined in Equation 3.6. The probability that the k^{th} plugin is the last in the chain is provided by the first row, whilst the probability of the first plugin in the chain is given by the first column. Finally, Equation 3.7 provides the formal definition for the Markov chain implemented, showing the nodes in the chain as a probabilistic series of states.

Using the transition matrix introduced in Equation 3.5, all the possible outcomes of the gathered dataset can be generated depending on their probabilities. The probabilities for each audio effect appearing at each position of the processing chain is illustrated in Equation 3.8.

$$P = \begin{bmatrix} 0.000 & 0.645 & 0.555 & 0.675 & 0.544 \\ 0.449 & 0.000 & 0.250 & 0.200 & 0.316 \\ 0.191 & 0.250 & 0.013 & 0.025 & 0.088 \\ 0.124 & 0.056 & 0.111 & 0.000 & 0.053 \\ 0.235 & 0.048 & 0.069 & 0.100 & 0.000 \end{bmatrix} \tag{3.8}$$

In this manner it is found that the most likely sequences of processing chains derived from the Markov model are: 1) EQ (29.0%), 2) reverb (12.8%), 3) compressor (10.6%), 4) distortion (8.3%), 5) compressor-EQ (6.2%) and 6) EQ-reverb (4.8%). These results further suggest that timbral transformations can be achieved through simple processing chains, usually without exceeding two processing modules. The fact that all base audio effects (i.e. single audio effect chains) appear in the first four positions regarding probable sequences illustrates that fact.

3.4 Processing chains

In total, 30 unique processing chains were constructed by the users in the duration of the experiment. It was also found that on average every processing chain was used 8.78 times. For this reason, chains that were implemented only once were considered outliers and were excluded from the following measurements. This action was performed as it is believed that there is not a sufficient number of applications of chains occurring once to evaluate their performance. This brought the number of unique chains to 19, and the total number of chains from 178 to 167.

The most popular chains in the study were: 1) EQ (27.5%), 2) reverb (12.5%), 3) compressor-EQ (11.9%), 4) distortion (8.9%), 5) EQ-compressor (8.9%) and 6) EQ-reverb (5.3%). The mean effect chain length was 1.64, suggesting that most chains in the dataset use one or two effects. The probability of an effect appearing in a given position in the chain is presented in Table 3.6. Considering that there were no restrictions on the number of plugins appearing in a chain, the majority of subjects used much shorter (1-3 effect) processing chains.

Effect	**1st**	**2nd**	**3rd**	**4th**
EQ (E)	0.44	0.43	0.21	0.33
Compressor (C)	0.22	0.28	0.25	0.33
Distortion (D)	0.15	0.10	0.16	0.00
Reverb (R)	0.17	0.18	0.38	0.33

Table 3.6: Probability of effects appearing per chain position.

For the first position, which includes chains of single effects, the EQ is the most popular module, appearing as 44% of the total instances, followed by the compressor at 22%, and the reverberation and distortion following with 17% and 15% respectively. This structure is retained in the second position, with the EQ still being the most popular, followed by the compressor and then the reverberation and distortion. However, this changes in the third position, where the most popular effect is the reverberation (38%), followed by the compressor (25%), EQ (20%) and finally the distortion (16%). The fourth position is split equally between the EQ, compressor and reverberation, all appearing 33% of the total instances, with the distortion never being used in that position.

3.4.1 Processing chain similarity

The similarity between chains can be measured through the use of MDS (Torgerson, 1952). A matrix of descriptor occurrences per chain is constructed, with a distance metric based on the co-existence of terms in each pairwise combination of chains, as detailed in Equation 3.9. Using this method, the distance between similar chains will tend towards 0, while the distance between chains that do not share any timbral adjectives will be 1. The resulting relationships can then be uncovered by implementing hierarchical, agglomerative clustering.

$$similarity = 1 - \frac{common\ terms}{total\ terms} \tag{3.9}$$

Figure 3.2 shows the results of hierarchical clustering on the MDS relationships between processing chains. In this instance, chains that are used to achieve similar terms, as is the case with EQ-compressor and compressor-EQ will display low cophenetic distances. This suggests that these combinations, even though they might be used with different parametrisations and for different instrument types and genres, are applied to achieve the same timbral transformations. Contrarily, plugin chains such as distortion and reverb, that do not share any semantic descriptors, will have high cophenetic distances.

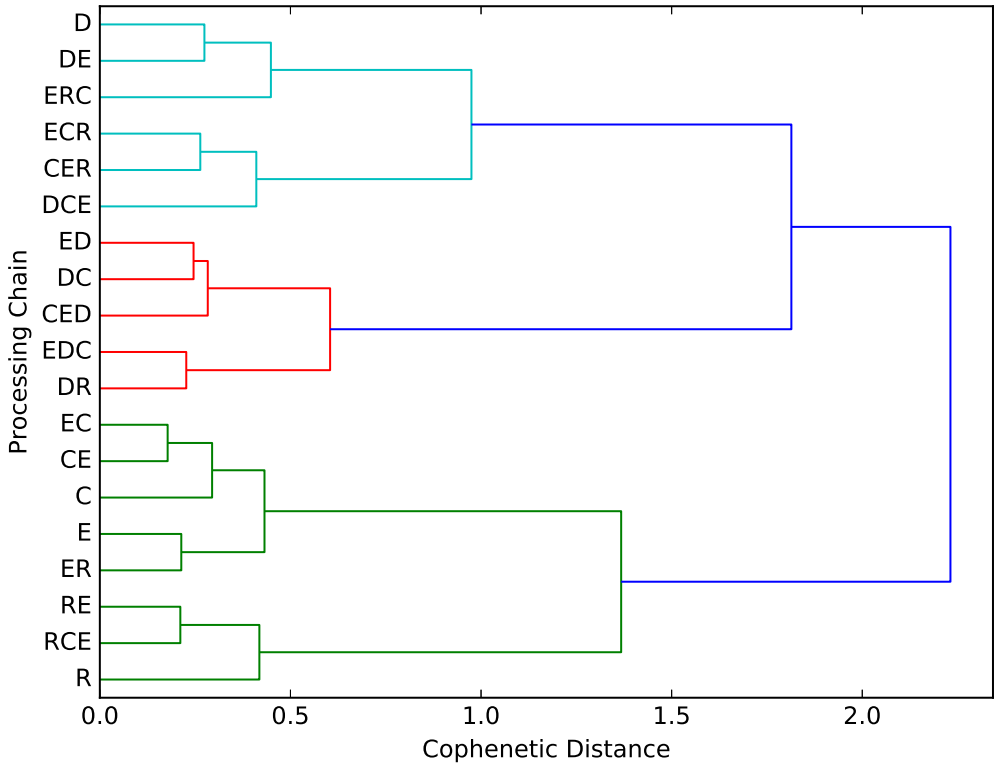


Figure 3.2: Hierarchical clustering of unique processing chains, comprised of combinations of equalisation (E), distortion (D), compression (C) and reverberation (R), based on term usage similarity.

3.4.2 Term similarity

In a similar manner to processing chains, the similarity of timbral adjectives can be assessed using the plugin chains selected to derive them. In this case, the descriptive terms can be represented through PCA, where each dimension defines the frequency by which a processing chain is implemented. Through this process a matrix D can be constructed, with $M \times N$ dimensions, where M represents the descriptors and N the unique chains. The $D(i, j)$ entry in the chain is the amount of times plugin chain j was used to achieve descriptor i. Similarly as in the processing chains, hierarchical clustering can then be used to uncover the relationships between different timbral adjectives.

The relationship between terms is presented in Figure 3.3. Here, it is found that terms are organised in three predominant groups: the first includes terms that mainly use distortion (*punch*, *fuzz*, *crunch*), the second includes adjectives associated with reverb (*room*, *dream*, *damp*) and finally a group with high generality is found, where terms are distributed across a range of plugin chains. The descriptor *warm* for example is achieved through 8 unique plugin chains, which make up for 42% of the unique chains in the dataset. All these interpretations of the same descriptor should be considered equally valid in being used for achieving a *warm* characteristic, and their existence shows how different routes can be used to perform the same timbral transformation. On the other hand, a term such as *fuzz* uses only 15% of the unique processing chains, a finding that suggests the adjective features a more narrow approach in the choice of audio effects.

3.4.3 Prevalent effect

To measure the prevalence of an effect in a processing chain the Euclidean distance of the audio features between the input and output stages is calculated. By comparing the distances between all nodes in the chain, the prevalent effect is defined as the node which caused the greatest change in audio features. This can be represented as a matrix P with dimensions $M \times E$, where M is the number of descriptors and E the number of audio effects. In addition, dimensionality reduction can then be applied to the matrix using PCA, which will project the semantic terms into a low-dimensional space. Finally, biplots can be adapted to the low-dimensional matrix to uncover the correlation of timbral adjectives with specific audio effects.

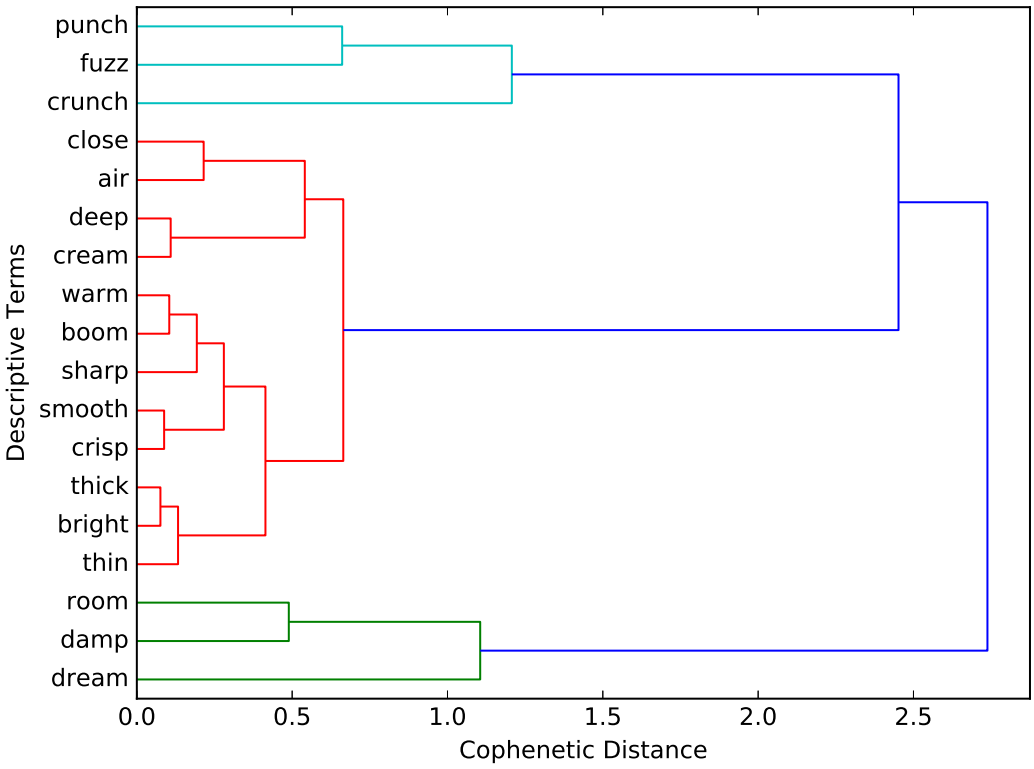


Figure 3.3: Hierarchical clustering of unique terms based on processing chain similarity.

Figure 3.4 displays the low-dimensional mapping based on the prevalent effect found in each timbral descriptor. From the biplot axes, it is shown that terms associated with a specific effect, in other words terms with low generality scores, are highly correlated with the effect dimensions. For example, *crunch* and *fuzz*, terms associated with distortion, display a correlation with the distortion dimension. Similarly, terms associated with the use of reverb, as *damp*, *dream* and *room* display a correlation with the reverb dimension, while terms associated with the compressor, in *sharp* and *boom*, also fall on the compressor dimension. On the other hand, terms that can be achieved through various audio effects, as in *warm*, are not positively correlated with any effect dimension.

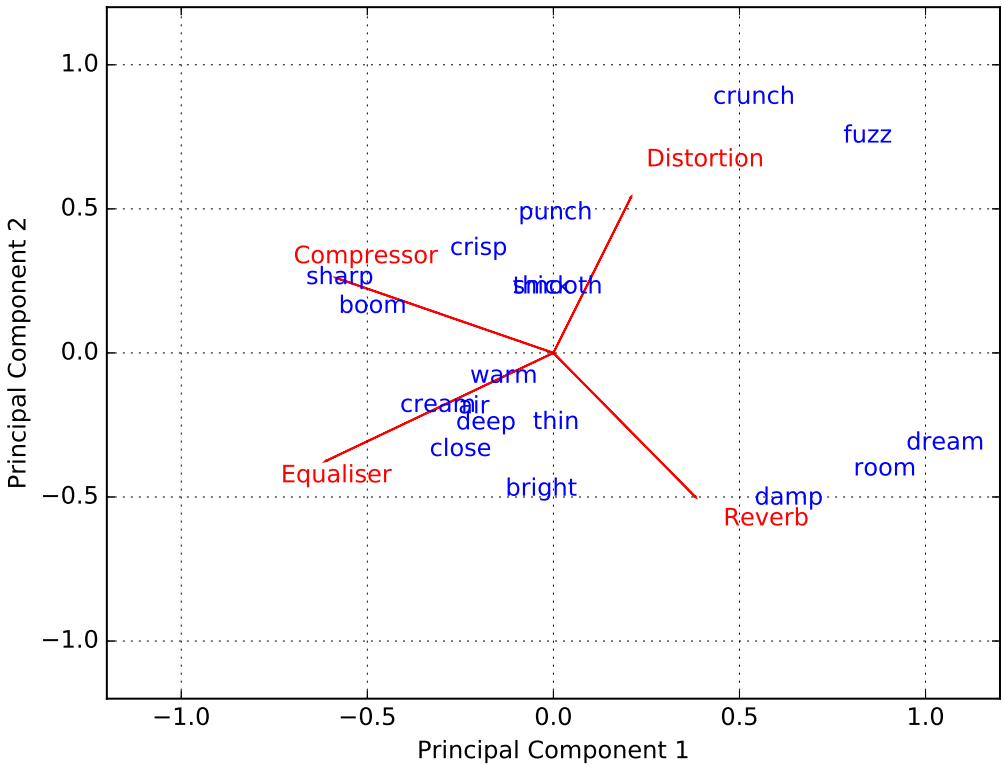


Figure 3.4: Low-dimensional semantic mapping (blue) with biplots computed using PCA, with prevalent audio effect dimensions (red) presented as biplots to highlight correlation of terms with individual audio effects.

3.5 Processing chain recommendations

Based on the analysis of single audio effect and processing chains, a recommendation system is proposed. This system will operate by providing recommendations of audio effects combinations for specific timbral adjectives.

Section 3.3.3 discussed the manner in which the processing chains gathered for this research can be treated as a Markov chain, where the probabilities of an effect appearing in the chain is associated with the previous state of the processing chain. By computing the state transition matrix (Equation 3.5) such a system can be constructed. In addition, it is possible to constrain the Markov chains based on the semantic terms used to derive them.

In this context a new transition state matrix P_d can be constructed for each descriptive term. By then generating the states of the Markov chain using P_d, chains with the highest likelihood in achieving descriptor d will be uncovered.

3.5.1 Unweighted recommender system

The Markov chain is able to predict the most common processing chains used for specific descriptors. For example, for the term *fuzz*, the model will generate a chain comprising a single distortion effect with a likelihood of 74.38%, and a combination of EQ-distortion with a likelihood of 16.53%. The probability matrices associated with each term can also be projected in a low-dimensional space, by implementing PCA, as is displayed in Figure 3.5. In this instance, terms that share plugin probabilities will be placed close together, as is the case with *room* and *damp*, or *sharp* and *punch*.

3.5.2 Weighted recommender system

In order to take advantage of the audio features information regarding the implementation of processing chains, an alternate approach is also considered for a weighted Markov chain. As was discussed in Section 3.4.3, the importance of an audio effect over the descriptor can be discovered by measuring the Euclidean distance between unprocessed and processed audio features. For this reason, an additional consideration should be the likelihood of a processing chain including a prevalent effect for achieving a specific descriptor.

Therefore, a weighting function is implemented in order to incorporate the importance of a prevalent effect appearing in the chain. Equations 3.10 to 3.12 display the process of deriving a weight w_p.

$$\mathrm{w}_p = \frac{\sum_{n=0}^{N-1} \sum_{l=0}^{L-1} f(\mathrm{d}, l) g(x(l), p)}{\sum_{n=0}^{N-1} n} \tag{3.10}$$

$$f(\mathrm{d}, l) = \begin{cases} 1, & \text{if } l = \text{argmax}(\mathrm{d}) \\ 0, & \text{otherwise} \end{cases} \tag{3.11}$$

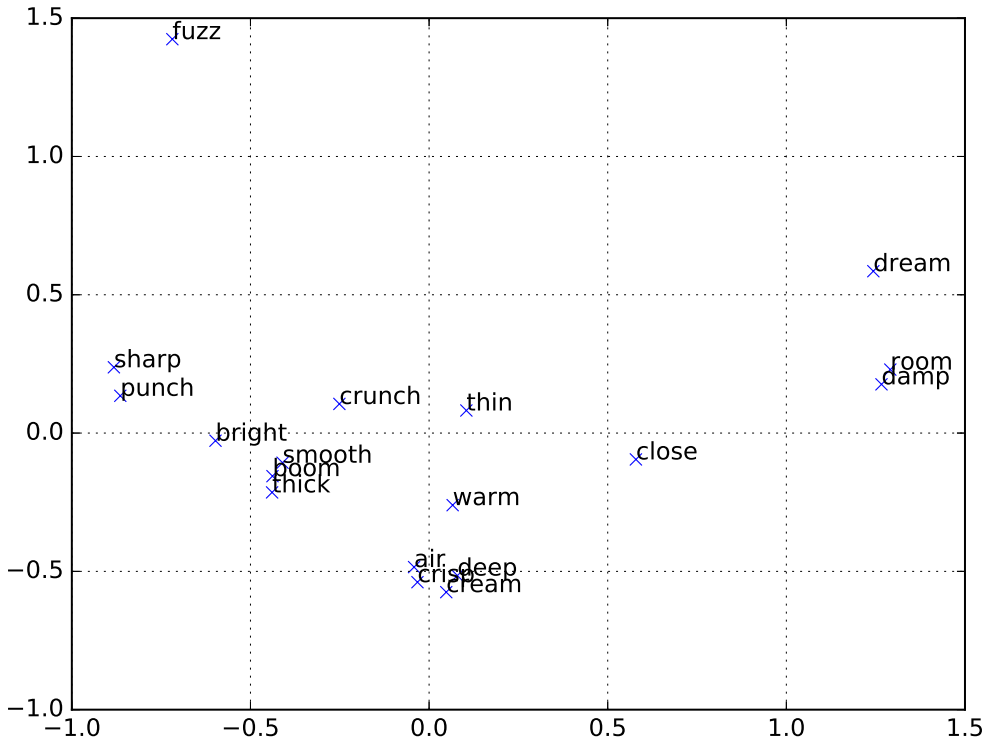


Figure 3.5: PCA low-dimensional projections of the unweighted Markov probabilities for all semantic terms.

$$g(x,p) = \begin{cases} 1, & \text{if } x = p \\ 0, & \text{otherwise} \end{cases} \tag{3.12}$$

Here, N is the number of chains, L the length of chain n, d a vector of the plugin Euclidean distances, as these were defined in Equation 3.3 and x a vector of plugin codes. Then the function f in Equation 3.11 will return a value of 1 if the plugin at position l is the most prevalent effect in the chain, while function g in Equation 3.12 returns a value of 1 if the plugin at position l is the same as plugin p. The weights are then multiplied with the corresponding row of the descriptor's transition matrix.

By weighting the probability of an effect appearing in a chain based on the prevalence

of each effect in achieving the specific descriptor, the original probability distributions of the Markov chain (Figure 3.5) can be altered. This process will not influence greatly terms that display low generality, such as *fuzz* and *dream*, which are associated with distortion and reverb respectively, but entries that encompass more general effects will display lower probabilistic scores.

For example, the transition state matrix of the unweighted Markov chain for the term *smooth* (P of *smooth*) will generate chains such as: EQ (16.67%), EQ-compressor (13.33%), reverb (11.11%) and distortion (11.11%). By implementing the weighting function, the new state transition matrix Pw of *smooth* will now generate the chains: compressor (28.87%), EQ (21.66%) and EQ-compressor (15.28%). The weighted probabilities distributions can also be projected in a low-dimensional plane through PCA, which is displayed in Figure 3.6.

3.5.3 Recommender system evaluation

To evaluate the performance of the processing chain recommender, its ability to retain the original structure between timbral adjectives is investigated. For this reason, structure preservation metrics can be implemented in order to measure the extent to which the structure is retained across spaces.

In dimensionality reduction, the rank-based metrics of *trustworthiness* and *continuity* (Venna and Kaski, 2005, 2006) are implemented to measure the accuracy of the transformed, low-dimensional space in preserving the structural information of the original, high-dimensional dataset. This is applied to the data, where the transformed space can be considered the Markov recommendation system. Equations 3.13 and 3.14 highlight the method in which these two metrics are derived, where the distances of the n entries in the two spaces (U and V) are converted to ranks (r and $\hat{r}$) between points i and j. The measurements then evaluate the distributions of datapoints in the respective spaces over a number of neighbouring datapoints (k).

$$T_k = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_i^{(k)}} (r(i,j) - k) \tag{3.13}$$

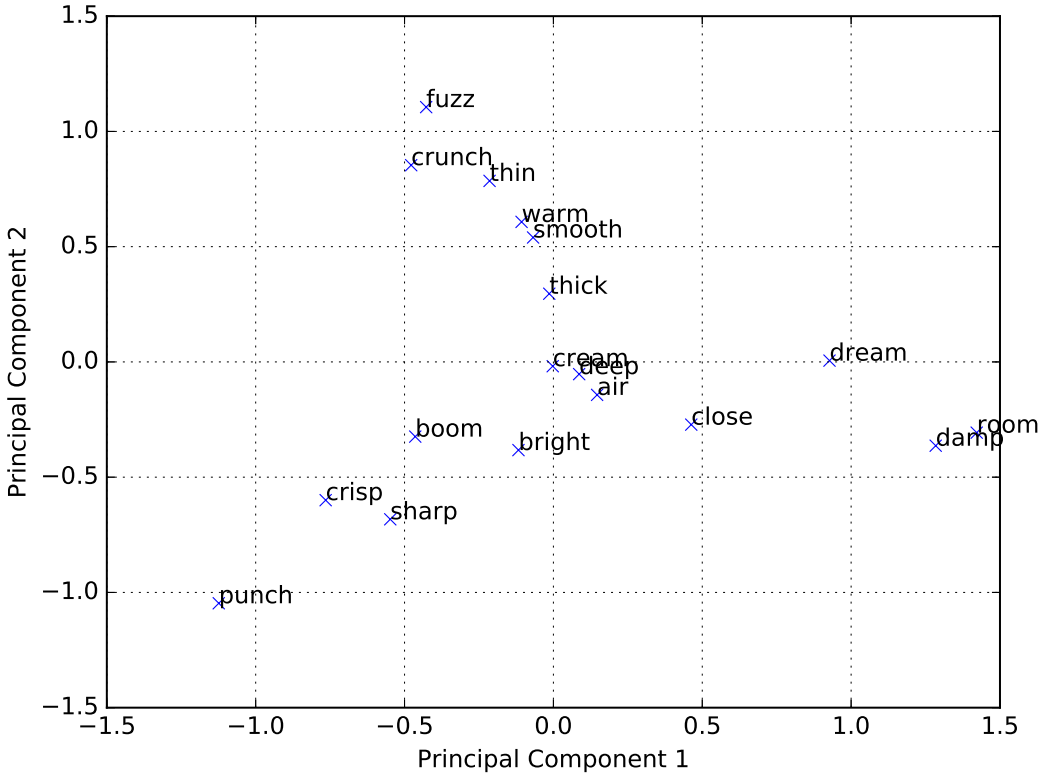


Figure 3.6: PCA low-dimensional projections of the weighted Markov probabilities for all semantic terms.

$$C_k = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in V_i^{(k)}} (\hat{r}(i,j) - k) \tag{3.14}$$

By measuring the structural preservation between the original descriptive term distribution (Figure 3.4) and the unweighted Markov chain probability distribution (Figure 3.5, it is found that the trustworthiness metric achieves a score of 0.78. The continuity metric performs with similar accuracy, achieving a score of 0.782.

When comparing the structures of the original descriptive term distribution (Figure 3.4) and the weighted Markov chain probability distribution (Figure 3.6), it is found that the trustworthiness metric achieves a score of 0.75, while the continuity score reaches 0.86.

3.6 Discussion

3.6.1 Processing chains based on semantic cues

A method for recommending processing chains based on individual semantic terms is presented. From Section 3.5.2, it can be shown that in the example of the term *smooth*, the weighted system was able to deduce that the compressor was the most important factor in achieving the term. In this instance, it is apparent that even though the compressor chain appears less often in the dataset, the impact that it has on achieving the *smooth* characteristic is more significant. Therefore, its presence should be highlighted in the generated chains. The system can therefore be considered informed of the impact of individual audio effects, and is capable of producing novel processing chains, that may not have been present in the original dataset.

Furthermore, the results for the trustworthiness and continuity suggest that there exists a direct connection between the relationships of timbral adjectives and the Markov model. It can also be deduced that the unweighted Markov space is capable of retaining the structure of the original space. However, it is shown that when the weighting function is implemented into the system, the preservation of relationships between timbral adjectives is more accurately represented.

This system can aid novice users by providing appropriate processing chains for timbral transformations. For example, the user will be able to provide an input descriptive term, for instance *fuzz*, and the system will produce chains that are most suitable for that descriptor. In the case of the term *fuzz*, those will include combinations of distortion and EQ.

3.6.2 The function of equalisation in processing chains

Effect appearance

Owsinski (2013) suggests that equalisation is the audio effect responsible for changing the timbre of an input sound, which is supported by prominent literature in the field (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005), where the majority of timbral adjectives are split into the respective frequency ranges used to derive them.

In the context of this research, the most popular effect implemented by the test subjects was the EQ, which makes up for 29% of the gathered processing chains. The next closest processing chain was the reverb, being used for 12.8% of the processing chains. In addition, considering there are 290 total plugin instances , the appearance of each effect in the context of the processing chains can be evaluated. The EQ appears 42.7% of the time within a processing chain. On the other hand the compressor is present 24.1%, the reverb achieves a value of 19.31% and finally the distortion appears 13.79% of the time. In this context the EQ is also the most frequently appearing component in any chain.

It is found that the EQ is also the most versatile processing unit, being used to achieve timbral transformations for every descriptor. In addition, the compressor is also shown to be a versatile processing module, used to derive all terms apart from *fuzz*. On the other hand, distortion is never applied for the terms *air*, *boom*, *cream*, *crisp* and *deep*. Similarly, reverb is never used for the terms *boom*, *punch* and *sharp*. Distortion and reverb can be therefore considered limited in regard to the descriptors they can represent.

With regard to the unique chains uncovered in this research, equalisation is found to be the most frequently used audio effect, being part of 73.6% of total chains, whilst the compressor appears in 57.78% of the unique processing chains. Less general effects, used for more specific processes, in distortion and reverb, both appear 42% of the time. This further suggests the importance of equalisation in altering the timbral characteristics of an input signal.

In addition it is found that EQ is the most frequently used effect in processing chains for 61.1% of the descriptors (Table 3.1). On the other hand, compression is implemented more often for the terms *boom* and *punch*, distortion for the adjectives *crunch* and *fuzz*, and finally reverb for the terms *damp*, *dream* and *room*.

Effect prevalence and salience

In measuring the prevalence of an effect, the EQ is found to be the more important element for achieving the timbral transformation in 30% of the processing chains (Figure 3.7). The compressor performs similarly, found to be the most important processing module for 29.21% of the processing chains. The reverb and the distortion are then found to be the most salient effects for 23.59% and 16.29% of the processing chains respectively. When considering the salience (Table 3.5) of the audio effects with regard to descriptors,

equalisation is found to be the most important element in achieving 38.8% of the descriptors, including *bright*, *close*, *cream*, *deep*, *sharp*, *thin* and *warm*. The compressor is found to be the prevalent effect for 33.3% of the timbral descriptors. Finally, effects that display a lower generality score, in distortion and reverb, are found to be prevalent in adjectives associated with their function, in *crunch* and *fuzz* for distortion and *damp*, *dream* and *room* for reverb.

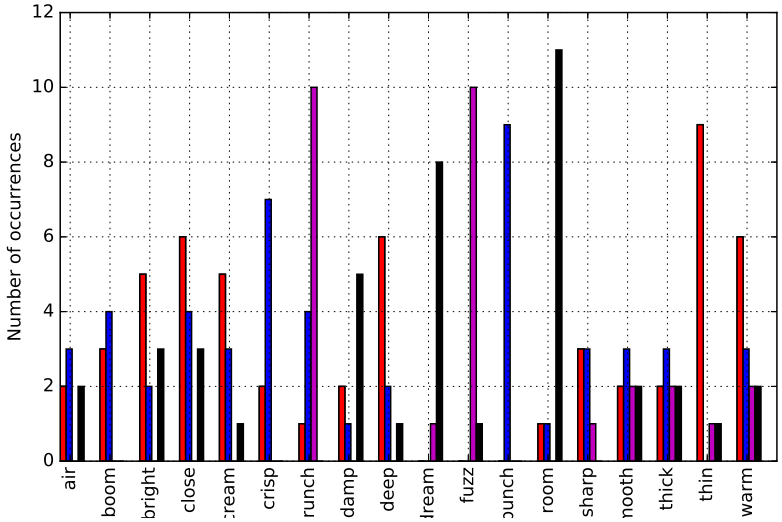


Figure 3.7: Prevalence of audio effect base types, EQ (red), compressor (blue), distortion (magenta) and reverb (black) for individual descriptors.

However, when equalisation appears in a processing chain, the effect that it has on the timbral transformation is less significant than the other effects in the chain. In the case of configurations such as CED, CER, CE, DCE, DE, ECR, EC, ED, ERC, ER, RCE and RE, the EQ is the least significant effect in the chain. The only instance where equalisation is not the least salient effect is in the EDC configuration, where it performs a slightly more significant task than the distortion module. This finding suggests that equalisation, when placed with other processors in a chain is used for fine-tuning or corrective purposes, rather than as the prominent effect. However, also in chains of single effects, changes imposed by equalisation do not correlate with the most significant changes in audio features. Non-linear effects, in distortion and compression, cause the most prominent changes in terms of Euclidean distance in features before and after processing, followed by equalisation and

finally reverb, as was illustrated in Table 3.5.

3.7 Conclusion

3.7.1 Overview

This chapter presents an analysis on the actions of sound engineers and music producers in achieving semantic transformations of sound. From the initial findings, an assessment is made regarding the role of external factors (i.e. genre, instrument, descriptor) in the choice of processing chain. As shown in Table 3.4, reverb and distortion achieve lower generality scores when measured across terms (0.464 and 0.291 respectively). These are the two lower scores in generality measurements, and they suggest that the two audio effects are used specifically for a subset of timbral adjectives in the dataset. This suggests that descriptive terms have a greater role in the choice of audio effects appearing in a chain, than the instrument and genre classes do. Additionally, it has been found that most processing chains used for semantic transformations are comprised of either one or two effects, suggesting that such operations do not require chains of great length. This is also expressed in the most frequently used processing chains, where the four most popular chains are the single effects (EQ, reverb, compression and distortion).

In addition, a method of assessing the similarity between processing chains and descriptive terms has been examined through the use of MDS and PCA, and also the relevance of the prevalent effect on the chain has been addressed. Through this information a system for processing chain recommendation has been proposed, that is capable of providing relevant processing chains dependant on an input term. By assessing the performance of the system through rank-based metrics it is found that the recommender can retain the structural information for both the unweighted and weighted versions, with the weighted version displaying an overall increase in the mapping accuracy.

Concerning the impact that individual effects have on the overall result, it has been found that non-linear processing modules in the distortion and compression can cause a bigger shift in the audio features (Table 3.5). This behaviour of non-linear audio effects can be attributed to their processing functionality. Both compression and distortion are capable of introducing artefacts in the magnitude spectrum of the sound they are processing, which

can cause a bigger change in the spectral and spectro-temporal audio features of an input sound than linear processing modules.

Furthermore, it has been found that distortion and reverberation are associated with specific terms, whereas compression and equalisation exhibit a higher generality across terms. However, the results have shown that equalisation is the more popular processing module for the majority of the descriptors presented in this study, something that further suggests that the audio effect is more directly associated with timbral transformations than compression.

As the majority of sound engineering literature suggests that the modification of the frequency components through equalisation is responsible for semantic timbral transformation (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005), this research has further examined this claim. Through this process it is found that equalisation is in fact the most frequently used chain, implemented as a single audio effect, and it has been used to achieve all the timbral adjectives presented in this study at least once. It is also shown that equalisation is the most frequently used effect within the context of a processing chain, suggesting that even when other audio effects are implemented, equalisation is still used in the majority of cases.

By considering the results of this research it becomes apparent that the sound engineering literature is correct in its claim that equalisation should be used for altering the timbral characteristic of an input signal. Since the analysis on processing chains has uncovered that essential role of equalisation in performing timbral modifications, this research will continue with the study of the audio effect in isolation. In this manner, detailed analysis can be performed regarding the parametrisation of equalisation to achieve specific timbral adjectives and the relationships between these terms.

3.7.2 Limitations

While this chapter has presented a novel method for comparing audio effect sequences through the use of semantic terminology, there still exist certain limitations to the study. Firstly, the research approach does not solve the problem of the order of audio effects in the processing chain, by which two audio effects in the chain (given that at least one

of them performs a non-linear process) cannot be swapped. While this study has found that chains of swapped audio effects, for example CE and EC, are used to achieve similar timbral adjectives, it cannot be considered that the effect that they have on an input sound is identical.

Furthermore, there exist certain limitations with regard to the processing chain recommendation system discussed in this chapter. While the recommender system provides sequences of audio effects that are appropriate to specific semantic terms, it does not provide automatic parametrisation of the audio effects in the chain. This is due to the high-dimensional nature of processing chains, by which the number of dimensions in the given dataset is comprised of all the parameters of all the audio effects in the chain. In addition, not only is the number of dimensions high, but the resulting dataset will also have a varying number of dimensions per processing chain. For example, a processing chain of CE has a different number of dimensions from a processing chain of CER. Moreover, an additional problem is the order of the dimensions. For example, while CE and EC have the same number of dimensions, their order is swapped and therefore a model cannot be used to predict appropriate parameters.

Chapter 4

Semantic equalisation

4.1 Introduction

Equalisation involves the application of multiple frequency dependent gains to an input audio signal. Applications for equalisation are found in all areas of sound production, from live audio engineering to recording, mixing and mastering. In music production, there are two forms of equalisation as Bazil (2009) suggests. Firstly, corrective equalisation, where problematic frequencies are identified and attenuated in order to prevent issues such as acoustic feedback, or masking of instruments inhabiting the same frequency range in a mix. The second type is creative equalisation, in which a user will perform a set of operations to change the timbral characteristic of a sound.

A subset of creative equalisation is the adjustment of EQ parameters to match a specific timbral adjective. This is often referred to as semantic equalisation, which is derived from the concept of timbre. As discussed in Chapter 2, timbral adjectives have been associated with the spectrum of an input sound from very early on (von Helmholtz, 1912; Lichte, 1941). Considering this aspect of timbre analysis, it is clear that as equalisation is able to adjust the spectrum of an input sound, it can influence its timbral characteristics. This is also found in existing sound engineering literature (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005) and on the findings presented in Chapter 3, where it was shown that equalisation is the most popular processing module out of four audio effects in achieving semantic terms. It is therefore necessary to evaluate the use of semantic equalisation in iso-

lation, in order to gain a deeper and more complete understanding of the actions performed by music producers and sound engineers during the modification of musical timbre.

In recent years there has been a series of systems attempting to uncover the underlying relationship between EQ parameters and timbral adjectives (Stables et al., 2014; Cartwright and Pardo, 2013; De Man, 2017). The methods for collecting data differ between most studies, which is something that provides divergences in both the data and the researchers' findings. For this reason, this chapter aims to present a comparative review of three existing datasets for semantic equalisation. By comparing the three datasets it can be understood whether agreement exists between the definitions of individual terms. Furthermore, it is possible to establish whether the structure between datasets shares similar characteristics, which would result in finding a set underlying structure between timbral adjectives. Moreover, the existence of confidence in regard to certain terms can be addressed.

Through the comparative review and the process of making the three different datasets comparable, global definitions regarding a shared terminology can be produced. This will allow the construction of a thesaurus of semantic terms, where synonyms and antonyms are discovered. This can potentially aid in deciphering the language of sound engineering. Therefore even when engineers use different words to describe sound, if those are synonymous then the process that they are describing can be considered almost identical. Finally, this chapter presents a dictionary of terms with corresponding frequency ranges, which is used to further decode the actions performed by a sound engineer. In this context for example, if an engineer is attempting to make something *bright*, they would most likely require an alteration of the high-end of the frequency spectrum.

4.2 Background

4.2.1 Equalisation

The concept of equalisation is said to have originated in the 1870s, where it was initially used for cancelling resonant frequencies in telephone communications, through the use of wave filters (Campbell, 1922; Zobel, 1923), or to flatten the frequency response of microphones (Bauer, 1962). The effect was then introduced on recording consoles in 1966, by Burgess Macneal and George Massenburg who subsequently published a paper on the

subject (Massenburg, 1972).

With the emergence of digital audio, EQs have been adapted from the analog to the digital domain through the use of digital signal processing techniques, allowing for more widespread use of the effect by novice users. Even though digital audio effects require the same level of expertise that analog equipment requires in order to be operated correctly, digital systems are less expensive, easily distributed and do not require physical space to be stored. In that way, mixing and mastering tasks in the digital domain require as little as a computer and speakers.

4.2.2 Analog filter design

Equalisation systems are composed of a series of filters, arranged either in cascade or in parallel. In this section some of the most commonly used filter designs will be discussed, as well as the advantages that each offers. Specifically, the designs that will be discussed are the Butterworth, Chebyshev (Type I and Type II), Bessel, Elliptic and Biquad filters.

Butterworth filter

The Butterworth filter, initially proposed by Butterworth (1930), was designed to have as flat a response as possible in the passband, eliminating any ripple (Smith et al., 1997). The transfer function for the filter is defined in Equation 4.1.

$$H(s) = \frac{1}{B_n(s)} \tag{4.1}$$

where n is the filter order and $B_n(s)$ the normalized Butterworth polynomials defined as (for n even and n odd respectively):

$$B_n(s) = \prod_{k=1}^{n/2} \left(s^2 - 2\cos\left(\frac{2k+n-1}{2n}\pi\right)s + 1 \right) \tag{4.2}$$

$$B_n(s) = (s+1)\prod_{k=1}^{n/2} \left(s^2 - 2\cos\left(\frac{2k+n-1}{2n}\pi\right)s + 1 \right) \tag{4.3}$$

It can be understood from this formula that the order of the filter directly corresponds

to the steepness of the transition band, i.e. the area between the filtered and unfiltered frequencies. A high filter order will produce a more steep transition band, while a lower filter order will increase the length of the transition band.

One of the key advantages of the Butterworth filter is that it provides a smooth, monotonically decreasing/increasing frequency response in the transition band, and it offers an optimal compromise between attenuation and phase response (Zumbahlen et al., 2011). Podder et al. (2014) concludes that this is due to the fact it does not feature any ripple in the passband and stopband sections. The frequency response of the filter is illustrated in Figure 4.1a.

Chebyshev filter

The Chebyshev filter design improves on the steepness of the roll-off from the Butterworh filter, but does so in the expense of passband (Type I) or stopband (Type II) ripple. The transfer function for Type I filter is shown in Equation 4.4.

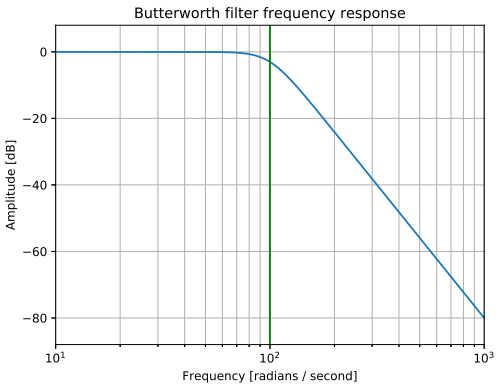
$$H(s) = \frac{1}{2^{n-1}\epsilon} \prod_{k=1}^{n} \frac{1}{s - s_k} \tag{4.4}$$

where n the filter order, ϵ the constant that defines the magnitude of the filter ripple in the pass band. Finally, s_k is defined as:

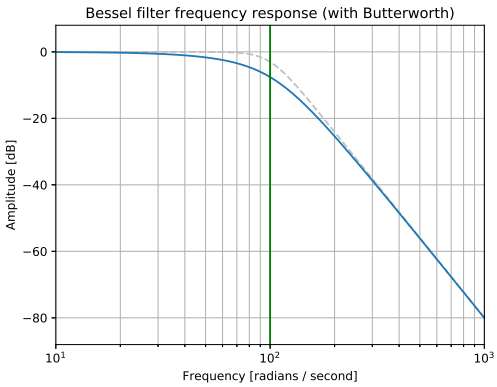
$$s_k = -sinh\left(\frac{asinh(1/\epsilon)}{n}\right) sin\left(\frac{\pi(2k-1)}{2n}\right) + jcosh\left(\frac{asinh(1/\epsilon)}{n}\right) cos\left(\frac{\pi(2k-1)}{2n}\right) \tag{4.5}$$

where $sinh$ and $cosh$ the hyperbolic sine and cosine functions respectively, and $asinh$ and $acosh$ the inverse functions.

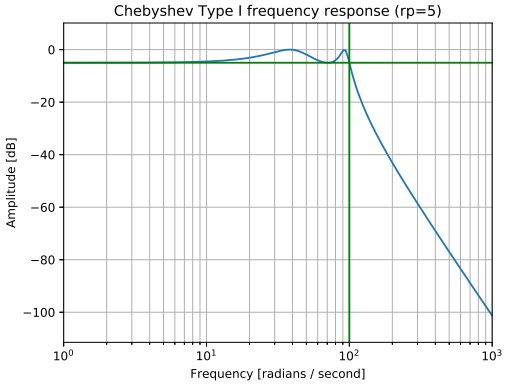
The Type II Chebyshev filter, known also as the inverse Chebyshev filter, does not feature any ripple in the passband, but displays ripple in the stopband section. It also requires more components to be constructed, and is therefore considered more computationally expensive than the Butterworth filter. The transfer function for the Type II filter is shown in Equation 4.6, and r_k is defined is defined in Equation 4.7.



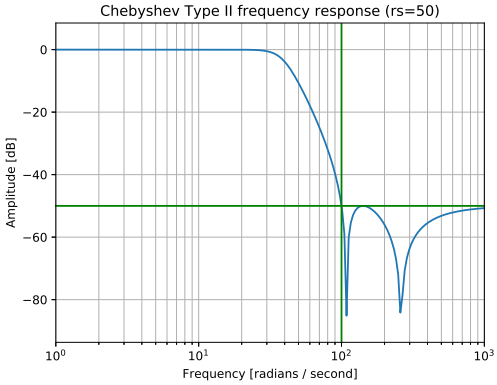
(a) Butterworth



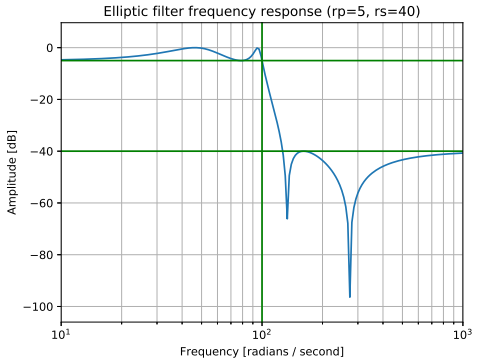
(b) Bessel



(c) Chebyshev Type I



(d) Chebyshev Type II



(e) Elliptic

Figure 4.1: Filter response for the Butterworth, Chebyshev (Type I and II), Bessel and Elliptic filter types.

$$H(s) = \epsilon \prod_{k=1}^{n} (s - r_k) \prod_{k=1}^{n} \frac{1}{s - s_k} \tag{4.6}$$

$$r_k = \frac{1}{-jcos\frac{\pi(2k-1)}{2n}} \tag{4.7}$$

Figures 4.1c and 4.1d display the differences for the Chebyshev filter design, where Type I (Figure 4.1c) features ripple in the passband, while Type II (Figure 4.1d) features ripple in the stopband.

Bessel filter

The Bessel filter was proposed by Thomson (1949), making use of Bessel functions to create a filter with maximally flat group delay, preserving the wave shape of the filtered signals in the passband. For this reason, and since it does not display any ripple in the stop band it can be considered similar to the Butterworth filter design. The transfer function of the filter is provided in Equation 4.8.

$$H(s) = \frac{c_{n,0}}{q_n(s)} \tag{4.8}$$

where n the filter order and q_n the reverse Bessel polynomials, defined in Equation 4.9 and the coefficients c defined in Equation 4.10.

$$q_n(s) = \sum_{k=0}^{n} c_{n,k} s^k \tag{4.9}$$

$$c_{n,k} = \frac{(2n-k)!}{2^{n-k} k!(n-k)!} \tag{4.10}$$

However, if the same filter order is used for Butterworth and Bessel filters, the stopband attenuation of the Bessel is lower than that of the Butterworth. In addition, as the filter attempts to produce a maximally flat transition band, the length of that transition tends to be wider than both Chebyshev and Butterworth filter designs. The difference between the Bessel and the Butterworth is illustrated in Figure 4.1b, where the Bessel filter response (blue line) features a prolonged transition band compared to the Butterworth (dashed line).

Elliptic filter

The elliptic filter is designed to have equal ripple (equiripple) in the stopband and passband sections. For this purpose it features a selectivity factor that can adjust the amount of ripple. This allows the filter to become more versatile and take on the characteristics of other filter designs. For example, when the ripple at both stopband and passband approach zero, the filter will become a Butterworth filter. On the other hand, adjusting the ripple independently on the passband and stopband sections will force the filter to become a Chebyshev filter (Type I or Type II).

The elliptic filter is generally defined by its gain function (Equation 4.11), rather than the Laplace transform.

$$G = \frac{1}{\sqrt{1 + \epsilon^2 R_n^2(\xi, \frac{\omega}{\omega_0})}} \tag{4.11}$$

where ω the angular frequency, ω_0 the cutoff frequency, ϵ the ripple effect and ξ the selectivity factor. Finally, R_n is the n order elliptic function, defined as:

$$R_n(\xi, x) = cd\left(n\frac{K(1/L_n)}{K(1/\xi)}cd^{-1}(x, 1/\xi), 1/L_n\right) \tag{4.12}$$

where cd the Jacobi elliptic cosine function, K the elliptic integral of the first kind and $L_n(\xi) = R_n(\xi, \xi)$ the discrimination factor.

An example of the filter's frequency response is provided in Figure 4.1e. where the ripple at the passband is set to a value 5 and the ripple at the stopband is set to a value 40. Here, the filter exhibits similar characteristics to the Chebyshev types presented in Figures 4.1c and 4.1d.

4.2.3 Digital filter design

The analog filter designs can be transferred to the digital domain, where the Laplace transform can be expressed as a z-transform. Highpass and lowpass filters have length of $n + 1$, while bandpass/bandstop filters have length $2n + 1$. For the analog filters the transfer function is shown in Equation 4.13, and it can be converted to Equation 4.14 in order for these to be represented as digital filters.

$$H(s) = \frac{B(s)}{A(s)} = \frac{b(1)s^n + b(2)s^{n-1} + ... + b(n+1)}{a(1)s^n + a(2)s^{n-1} + ... + a(n+1)} \tag{4.13}$$

$$H(z) = \frac{B(z)}{A(z)} = \frac{b(1) + b(2)z^{-1} + ... + b(n+1)z^{-n}}{a(1) + a(2)z^{-1} + ... + a(n+1)z^{-n}} \tag{4.14}$$

Biquad filter

As the computational cost of digital filters is dependant on the complexity of the filter design, some of the aforementioned models can be considered expensive. For this reason, filters such as the biquad have been popular in digital implementations.

The biquad filter (Tow, 1968; Thomas, 1971) implements a transfer function that is the ratio of two quadratic equations (Equation 4.15). This filter type provides flexibility in easily representing different filter types (*low-pass*, *high-pass*, *low-shelf*, *high-shelf*, *band-pass*, *peak/notch*), determined by the variables ($b_0, b_1, b_2, a_0, a_1, a_2$). The coefficients of the biquad are often normalized so that $a_0 = 1$ depicted in Equation 4.16.

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2}} \tag{4.15}$$

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{4.16}$$

4.2.4 Filter arrangement

To increase versatility for filter implementations, individual filters are combined. In this manner an array of filters (filterbank) is created. This allows sound engineers to control a number of different filters through altering the parameters of a single processing module, which is known as the EQ.

There are two methods of arranging filters into EQs, either by series (cascade) or parallel combination. This is performed by combining the transfer functions of the individual filters, either by multiplying (series method) or summing (parallel method) these together.

Cascade

Figure 4.2 displays the design of a cascade filterbank. In this case the input signal $x(n)$ is passed through the first filter transfer function ($H_1(z)$) to produce filtered signal $x'(n)$. The new signal is then processed through the second filter transfer function ($H_2(z)$) to produce the final output of the EQ, signal $y(n)$.

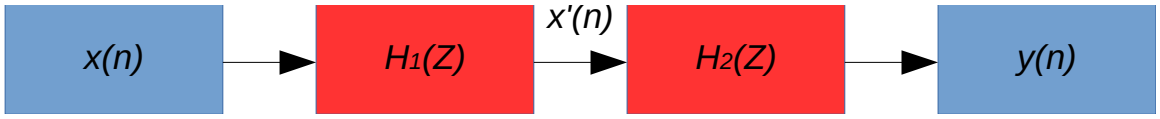


Figure 4.2: Cascade EQ design structure. Here the filters are placed in series with the output of the previous filter, being passed to the next. Blue boxes denote the input and output stages and red boxes depict the filters of the system.

The process is expressed in Equation 4.17, where $H_{eq}(z)$ the final transfer function of the EQ, M the total number of filters, m the current filter and $H_m(z)$ the current filter's transfer function. The formula iterates through the transfer functions of all M filters, multiplying these together.

$$H_{eq}(z) = \prod_{m=1}^{M} H_m(z) \tag{4.17}$$

Parallel

Figure 4.3 displays the design of a parallel filterbank. Contrarily to the cascade filter design, the transfer function of parallel systems is the product of the functions of individual filters. The input signal $x(n)$ passes through both $H_1(z)$ and $H_2(z)$ transfer functions to produce two new signals $x'_1(n)$ and $x'_2(n)$ respectively. These outputs are then summed to produce the final signal $y(n)$.

This process is expressed in Equation 4.18 where, as in Equation 4.17, $H_{eq}(z)$ the final transfer function of the EQ, M the total number of filters, m the current filter and $H_m(z)$ the current filter's transfer function. However, in this case there is an additional control over the gain parameter of filter m, presented as G_m.

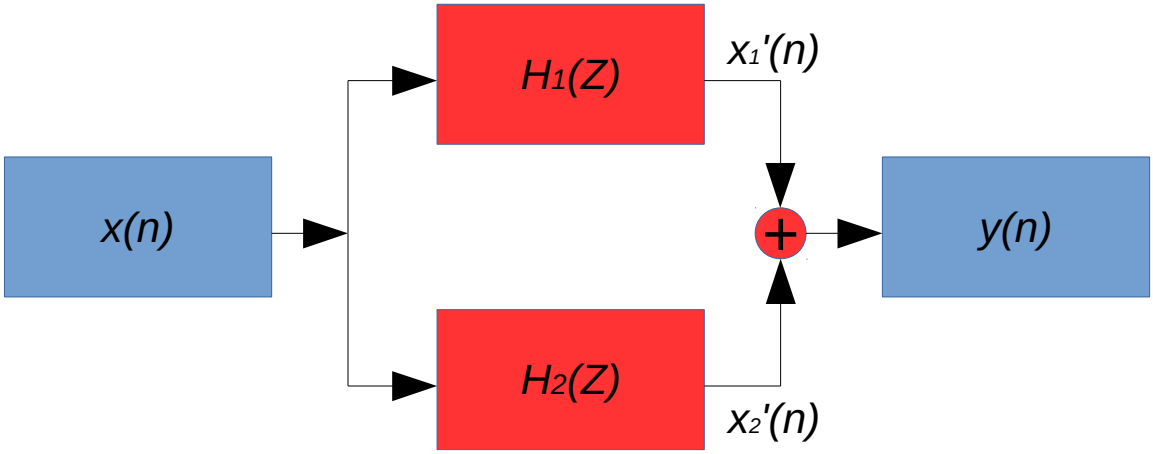


Figure 4.3: Parallel EQ design structure. Here the filters are placed in parallel and their individual outputs are summed. Blue boxes denote the input and output stages and red boxes depict the filters of the system.

$$H_{eq}(z) = \sum_{m=1}^{M} G_m H_m \tag{4.18}$$

4.2.5 Filter types

The most common filter types include: *low-pass*, *high-pass*, *bandpass* and *bandstop*. In the case of EQ implementations, the types of filters most frequently applied include shelving filters, for the lowest and highest frequency bands of the EQ, and peak filters.

Shelving Filters

For shelving filters, a gain G is applied to all frequencies above or below the centre frequency (ω_c), in order to attenuate, or boost, the high, or low, frequency components.

Equation 4.19 illustrates the function for a first-order low-shelf filter, where due to the choice of gain at the centre frequency (ω_c), the effect of the filter will be symmetrical for boost ($G > 1$) and cut ($0 < G < 1$) implementations. This means that for a boost of value G and a cut of value $1/G$ the filter will produce a mirrored response on the magnitude spectrum around the Nyquist frequency.

The first-order low-shelf filter formula (Equation 4.19) can then be adapted in order

to be implemented as a high-shelf filter. This is performed by replacing the gain value G with $1/G$, which will invert the filter. Then the transfer function is multiplied by G, which will shift the magnitude response vertically so that the gain G is achieved at the higher frequencies. Finally, the numerator and denominator of the transfer function are multiplied by $\sqrt{G}$ in order to cancel any divisions by G. The resulting transfer function is presented in Equation 4.20.

$$H_{LS}(z) = \frac{Gtan(\omega_c/2) + \sqrt{G} + [Gtan(\omega_c/2) - \sqrt{G}]z^{-1}}{tan(\omega_c/2) + \sqrt{G} + [tan(\omega_c/2) - \sqrt{G}]z^{-1}} \tag{4.19}$$

$$H_{HS}(z) = \frac{\sqrt{G}tan(\omega_c/2) + G + [\sqrt{G}tan(\omega_c/2) - G]z^{-1}}{\sqrt{G}tan(\omega_c/2) + 1 + [\sqrt{G}tan(\omega_c/2) - 1]z^{-1}} \tag{4.20}$$

where G the gain applied to the filter and ω_c the centre frequency in radians ($0 \leq \omega_c \leq \pi$).

Peak Filters

Apart from the shelving filters, the remaining sections of an EQ comprise peak, or notch, filters. These are usually second-order filters that provide a magnitude response which has applied gain G and width specified by the filter bandwidth or Q-factor around the center frequency ω_c, while the unaffected frequencies will all be set to unity gain. The transfer function for the second order peak/notch filter is presented in Equation 4.21, where the bandwidth B can be replaced by the Q-factor, as depicted in Equation 4.22.

$$H_{PN}(z) = \frac{\sqrt{G} + Gtan(B/2) - [2\sqrt{G}cos(\omega_c)]z^{-1} + [\sqrt{G} - Gtan(B/2)]z^{-2}}{\sqrt{G} + tan(B/2) - [2\sqrt{G}cos(\omega_c)]z^{-1} + [\sqrt{G} - tan(B/2)]z^{-2}} \tag{4.21}$$

$$Q = \frac{\omega_c}{B} \tag{4.22}$$

4.2.6 Equalisation interfaces

The two main interfaces of equalisation used today are parametric equalisation and graphic equalisation, both of which are capable of adjusting the spectral envelope of a signal (Read,

1952), and can therefore be considered tools for the modification of timbre.

Parametric equalisation

The first use of the term *parametric* appeared in a paper presented by Massenburg (1972) regarding a novel equalisation design. Parametric EQs are usually comprised of two shelving filters (for the low and high ends) and an array of peak filters between the shelving filters. The filters of the EQ are always arranged in cascade, where the effect of each filter is summed so that the overall effect of the EQ is cumulative on the decibel scale.

Parametric EQs allow for the modification of the gain, centre frequency and bandwidth (or Q-factor) of the midrange bands (using peak filters), while also featuring the capability of adjusting the gain and cutoff frequency of the shelving filters. For this reason the parametric architecture is considered to be the most powerful and flexible application of equalisation.

The main advantage of parametric EQ is that they expose the cutoff frequency and bandwidth of filters to the users. In that way, it is possible to add a peak or notch filter at any point in the frequency domain, without limitations. This can be performed in order to either amplify a prominent frequency range of an individual track, resulting in it being perceived clearer in the overall mix, or in order to attenuate any problematic frequencies and remove unwanted sounds, such as humming noises (Brandt and Bitzer, 2012) or acoustic feedback (Van Waterschoot and Moonen, 2011). Both peak filters and shelving filters display the same degree of freedom regarding their placement in the frequency spectrum.

The representation of parametric EQs in the analog world can be considered complicated, as it does not include a visualisation of the frequency spectrum. Figure 4.4 provides an example of analog parametric EQs, where the user will not have a graphical interface that will provide a visualisation of the filters' frequency responses.

On the other hand, when parametric EQs are implemented in the digital domain their operation becomes much clearer through the Graphical User Interface (GUI). Figure 4.5 shows an example digital parametric EQ, where the main functions of the system are displayed alongside a GUI, which can both illustrate the filter parameters and also alter them. Because of its flexibility and the possibility of frequency spectrum representation, the parametric EQ is the most popular equalisation design used in digital audio effects.

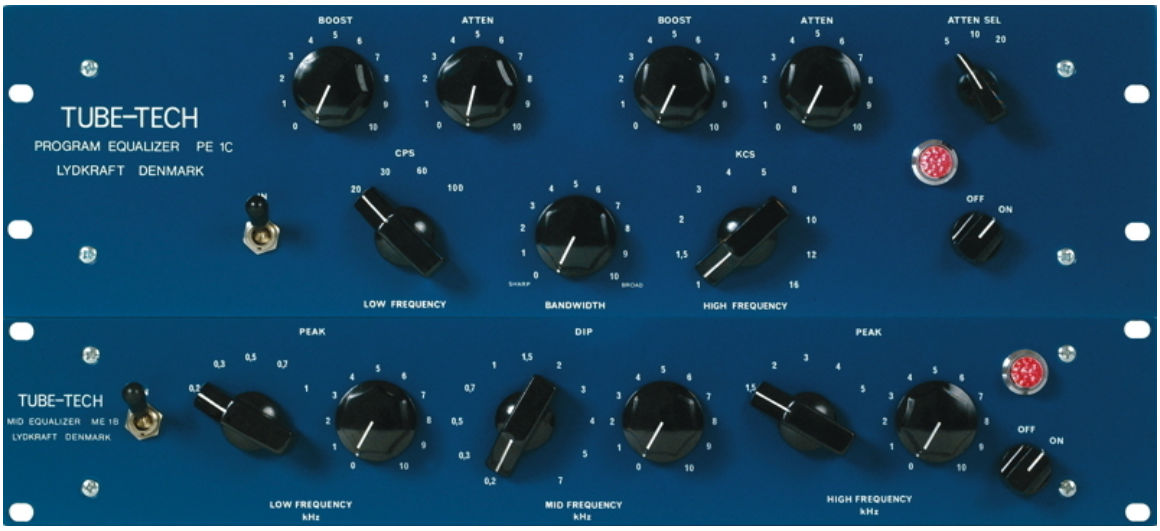


Figure 4.4: The Tube-Tech PE 1C and ME 1B analog parametric EQs.

Graphic equalisation

Graphic equalisation is a simpler and less flexible implementation than parametric equalisation. Whereas the parametric interface allows for modification over the gain, cutoff frequency and Q-factor of each individual filter, the graphic interface allows for modification only of the gain of individual filters, while the centre frequency and the Q-factor parameters remain static. For this reason, this architecture might lack the flexibility of parametric EQs but it is a simpler and easier system to use.

Graphic EQs can be implemented by using a cascade of filters (Hirata, 1981; Greiner and Schoessow, 1983; Abel and Berners, 2004), where each filter gain determines the overall magnitude response of the EQ, independently of each of the other bands. Additionally, graphic EQs can use parallel banks of bandpass filters (Greiner and Schoessow, 1983; Tassart, 2013; Ramo et al., 2014), where each bandpass filter produces a gain based on the gain slider position. In order to avoid interaction between different filters, the magnitude response of each bandpass filter should be approximately zero at all other frequencies.

Figure 4.6 shows an analog graphic EQ interface, where the resulting frequency spectrum can be derived from the position of the individual filter gain values. Here, the analog graphic EQ can be considered to have a simpler interface to the analog parametric EQ, due to its graphical representation of the frequency spectrum. For this same reason the imple-

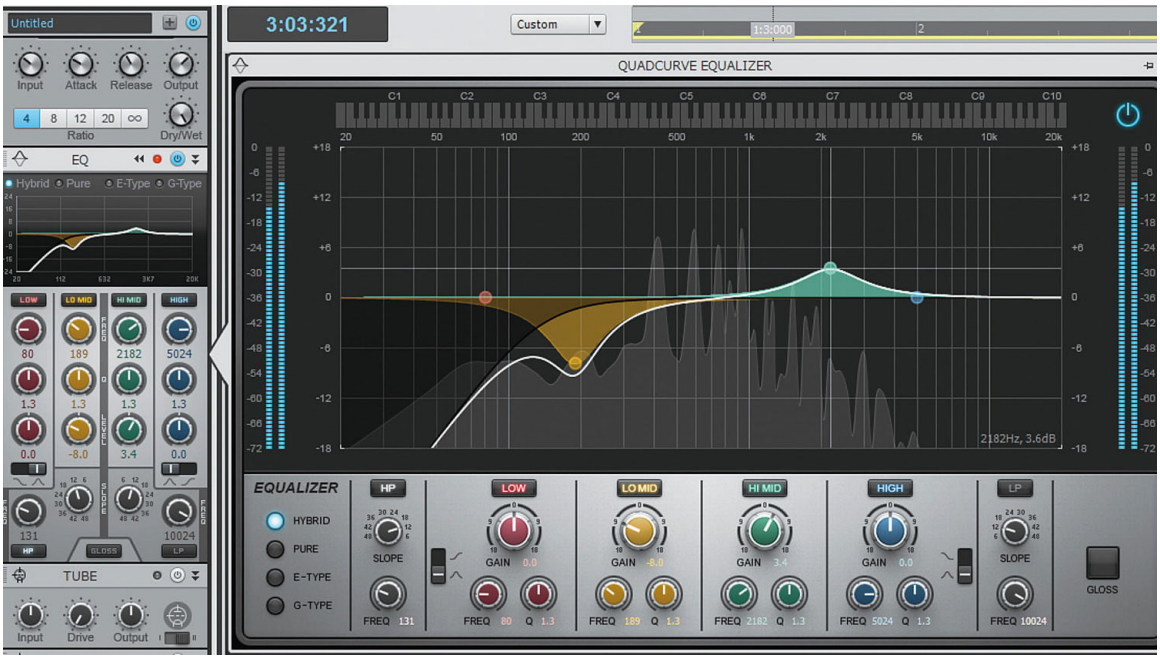


Figure 4.5: The Cakewalk Quadcurve digital parametric EQ.

mentation of graphic EQs into the digital domain does not provide any further advantages to the parametric implementation.

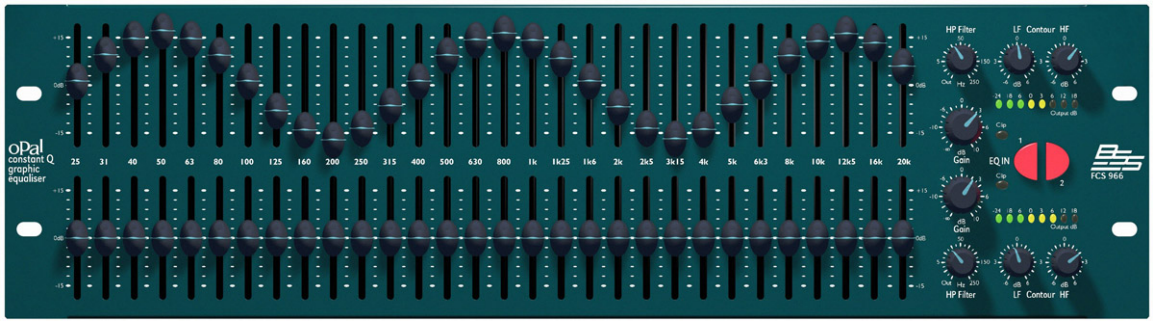


Figure 4.6: The BSS FCS966 Stereo Graphic analog equalizer.

While graphic equalisation features a more limited design regarding the functionality of the system, it is still widely used in audio effects. An example is the use of stereo graphic EQs on the PA output in live sound events, where they offer an advantage in "tuning" the room modes of the venue.

4.2.7 Term definitions

In the context of equalisation, sound engineers will make use of timbral adjectives in order to describe their choice of parameter settings. For instance, engineers and producers tend to boost high frequencies of the input sound in order to *add air*, or they might apply gain to the lower mid frequencies to make something *warmer*. Even though the parametrisation has some dependence on the type of input audio being processed, there seems to exist a common heuristic basis for achieving certain terms. Numerous texts by professional music producers include examples of the association between frequency components and descriptive terminology (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005).

Various EQ studies, presented in Section 4.3.1, investigate the connection between timbral adjectives and parameter settings either through collection of heuristic guidelines (De Man, 2017), or through crowd-sourcing systems that collect data from a wider audience. The SocialFX (Zheng et al., 2016) and SAFE (Stables et al., 2014) datasets are both investigative systems, attempting to uncover the relationships between audio effect parametrisation and timbral transformations.

By comparing the entries from the three datasets, SAFE, Social and Heuristic, a comparative review can be performed. In this study, the aim is to uncover the agreement between term definitions across different datasets, the relationships exhibited between terms (synonymous/antonymous terminology), and the correspondence of terms to frequency ranges.

4.3 Methodology

4.3.1 Datasets

SAFE EQ

The SAFE EQ is one of the four audio effects included in the SAFE project (Stables et al., 2014), which aims to investigate the role of audio effects in altering the timbral characteristics of a sound. The SAFE EQ operates in a DAW environment - as is the case with most standard digital audio effects - with the added functionality of storing the user's settings

and extracting the audio features of the track being processed.

The SAFE EQ uses a parametric design, which includes five biquad filters arranged in series and consists of one low-shelving filter (LS), one high-shelving filter (HS) and three peaking filters (P). The peaking filters can be parametrised in terms of their centre frequency, gain and Q-factor, while HS and LS are parametrised by their centre frequency and gain. The ranges of all filter parameters are displayed in Table 4.1.

n	**Assignment**	**Range**	**n**	**Assignment**	**Range**
0	LS gain	−12–12 dB	7	P_1 Q	0.1–10
1	LS Freq	22–1000 Hz	8	P_2 Gain	−12–12 dB
2	P_0 Gain	−12–12 dB	9	P_2 Freq	220–10,000 Hz
3	P_0 Freq	82–3900 Hz	10	P_2 Q	0.1–10
4	P_0 Q	0.1–10	11	HS Gain	−12–12 dB
5	P_1 Gain	−12–12 dB	12	HS Freq	580–20,000 Hz
6	P_1 Freq	180–4700 Hz			

Table 4.1: The SAFE EQ interface range of parameters for its five bands, allowing for control over the gain and centre frequency of the low-pass and high-pass filters, and over the gain, centre frequency and Q factor over the three mid-band filters.

The SAFE EQ dataset contains 1700 saved terms [1], 248 of which are unique. From the 248 unique terms, 71 are presented multiple times in the dataset. The ten most frequent terms of the dataset (appearing most times) are *warm* (535), *bright* (514), *clear* (18), *present* (14), *air* (12), *thin* (12), *clean* (11), *bass* (10), *boom* (9) and *full* (8).

SocialEQ

Social EQ (Cartwright and Pardo, 2013) is a sub-division of the SocialFX project, investigating the role of equalisation in changing the timbre of a sound. The application is implemented as a 40-band graphic EQ, which is capable of running in a web browser. The subjects visiting the website are asked to initially choose a term, which the system will then try to implement through changing the parameters of the graphic EQ. Audio samples with different equalisation parameters are then displayed back to the user, and the subjects are asked to rate the extent to which each sample has been able to achieve the specific timbral

[1] Accessed 15 August 2017

transformation. The implementation of SocialEQ comes with a number of pre-defined audio samples that the users can choose from, with the additional possibility of uploading a sample of their choice.

The filter design implemented for the SocialEQ application is the squared Butterworth filter, also known as the Linkwitz-Riley (L-R) filter (Linkwitz, 1976). The L-R filter is constructed by placing two Butterworth filters (a *low-pass* and a *high-pass*) in cascade at the centre frequency. This architecture provides an advantage over the Butterworth filter as the gain at the cutoff frequency will be 0 dB, where summed Butterworth filters would achieve a gain of +3 dB. The filters are then cascaded to form a 40-band filterbank. The system then learns an EQ curve with relative gain values, ranging from -1 to +1. However, the users are able to change that to a range up to -20 to +20.

SocialEQ contains a total of 731 entries [2], 295 of which are unique terms. From the 295 unique terms, 92 are presented mutliple times in the dataset. The ten most frequent terms of the dataset (appearing most times) are *warm* (58), *cold* (25), *soft* (24), *loud* (22), *happy* (19), *bright* (16), *soothing* (15), *harsh* (15), *heavy* (12) and *cool* (11).

Heuristic EQ

The existing literature of sound engineering practices (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005) expands on the role of semantic terms in music production. De Man (2017) used heuristic guidelines from the existing literature and combined them to form a large corpus of timbral adjectives and their corresponding frequency ranges (a sample was provided in Table 2.4). In total 95 terms are identified, and 210 interpretations are found for all configurations detailed in the list. These heuristic guidelines can be used to generate EQ curves. The ten most frequent terms of the dataset (appearing most times) are *air* (9), *muddy* (7), *warm* (6), *boom* (6), *present* (6), *sibilant* (6), *body* (5), *fat* (5), *full* (5) and *clear* (5).

[2] Accessed 15 August 2017

4.3.2 Comparison of datasets

All three datasets appearing in this study use different architectures for performing equalisation. In order to compare these, adjustments need to be made to create a uniform format for them. The Heuristic EQ dataset will need to be transformed from guidelines to full EQ curves. By providing the range of the frequencies, in maximum (F_{max}) and minimum (F_{min}) values, required to modify the timbre, the EQ curves will need to encompass that range in a fashion similar to applying a peak filter over that region. For this purpose, filters corresponding to these values can be designed. The necessary elements of this filter include the centre frequency (F_c), the Q-Factor (Q) and a gain value.

Starting with the frequency information provided by these ranges, the centre frequency (F_c) of a filter can be calculated, as in Equation 4.23:

$$F_c = (F_{max} + F_{min})/2 \tag{4.23}$$

The bandwidth of the filter can be found, as in Equation 4.24, which in turn will provide an approximation of the Q-factor parameter (Equation 4.25) needed to appropriately fit the frequency range to the filter.

$$BW = F_{max} - F_{min} \tag{4.24}$$

$$Q = F_c/BW \tag{4.25}$$

As no guidance is provided regarding the gain parameter, other than whether a boost or a cut of the frequency region is required, the value can be set arbitrarily. For this reason, if the timbral adjective dictates a boost over the frequency range, then a gain of +3 dB is applied, while if a cut is demanded, the gain is set to -3 dB. In this manner, the heuristic definitions will be consistent.

Finally, since some terms in the corpus of descriptors contain more than one definition, the mean EQ curve will need to be calculated. This is a trivial task, in which the frequency responses are added together and subsequently normalised in order to get the final representation of the term. Figure 4.7 depicts four EQ curves generated through the aforementioned process.

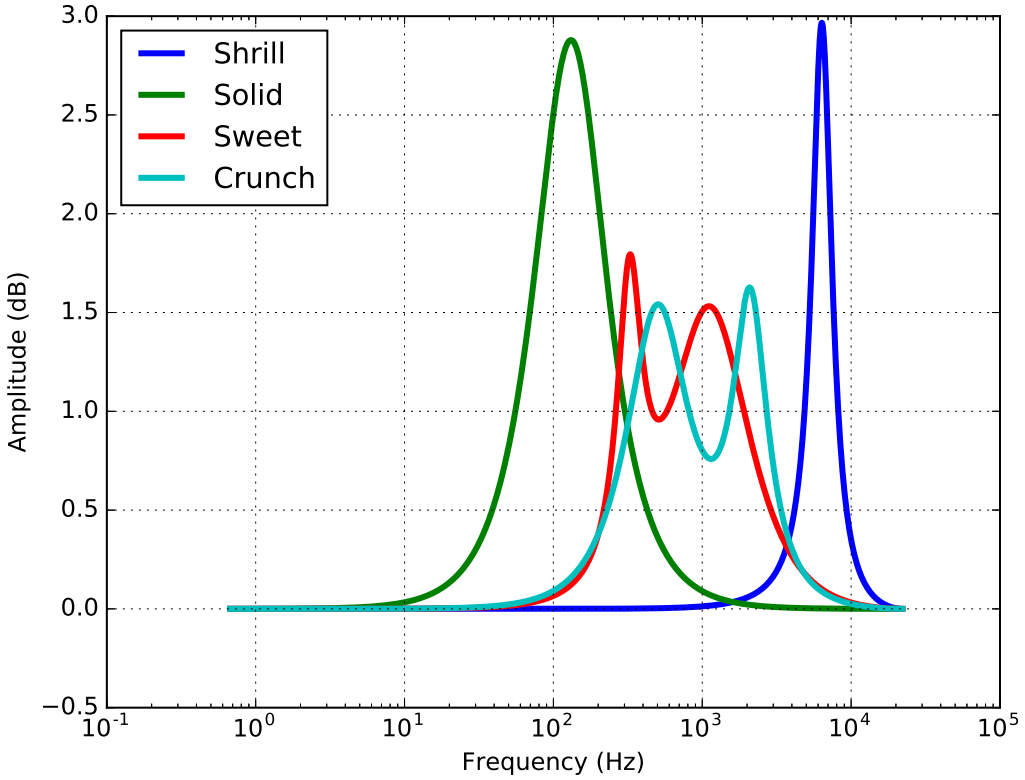


Figure 4.7: Constructed mean frequency distributions for the heuristic definitions of *shrill, solid, sweet* and *crunch*.

By interpreting the heuristic guidelines into EQ curves, the Heuristic EQ dataset becomes comparable with the SAFE EQ dataset. However, as the SocialEQ dataset does not feature a parametric EQ architecture, but a 40-band graphic EQ, one further adjustment needs to be made in order to perform the comparative analysis. As the SocialEQ system defines specific frequency bins for each band, it is possible to sample the corresponding 40 bins from both SAFE EQ and Heuristic EQ. In this manner there is no need to compare the parameters of the systems but rather their corresponding EQ curves.

Apart from differences regarding the architectural implementation of equalisation between datasets, the process of gathering data is also substantially different. As SAFE EQ operates within a DAW, it can be thought that individuals using such a system will be well versed in music production techniques. On the other hand, SocialEQ uses Amazon

Mechanical Turk for finding test subjects, which suggests that a broader audience has participated in the given task. Finally, Heuristic EQ represents the guidelines provided by renowned sound engineers.

4.3.3 Comparative review

Between all datasets it is found that 529 unique terms are included. Considering the number of unique terms in each group, the maximum possible percentage of common terms is that of the lowest number in a dataset, in this case 95 for Heuristic EQ. It is found that 29 terms (5.48%) are common between all three datasets. This low value is likely due to the fact that SocialEQ and SAFE EQ do not provide any limitations to the type of words/phrases subjects use to describe timbral transformations.

Dataset	*Heuristic*	*Social*	*SAFE*
Heuristic	100	10.48	16.27
Social	10.48	100	10.81
SAFE	16.27	10.81	100

Table 4.2: Percentage of common terms between the SAFE, Social and Heuristic datasets.

Table 4.2 displays the overlap between individual studies, where it is found that the SAFE and Heuristic EQ datasets share most terms (16.27 %). The overlap between SocialEQ and Heuristic EQ reaches a percentage of 10.48%, while SAFE EQ and SocialEQ share 10.81% of terms. This can be attributed to the data gathering process implemented by different systems. SocialEQ acquired information through Amazon Mechanical Turk, while SAFE EQ operates within a DAW environment. The fact that SAFE EQ is built to operate as a typical audio plugin guarantees that subjects will have experience with the basic concepts of sound engineering, something that is not necessarily applicable in SocialEQ. From the popular terms in SocialEQ, it can be determined that a number of these do not correspond to the terminology typically used by sound engineers. Terms such as *cold, happy* and *cool* are not usually attributed to audio effects parametrisation. In the cases of terms such as *happy* and *pleasant*, both of which appear in the SocialEQ dataset, it can be argued that they do not appear in the appropriate category of semantic terminology (presented in Section 2.2.1), as those are associated with the *indexical* syntax category (instead of the iconic syntax category). This is a result of differences in the data gathering process, which

does not ensure a baseline regarding the subjects' backgrounds and knowledge of sound engineering. Another interesting finding is that the two bigger datasets (Social and SAFE) share only 10.81% of terms. This high number of unique terms provides an indication as to the range of possible terminologies for describing timbral transformations.

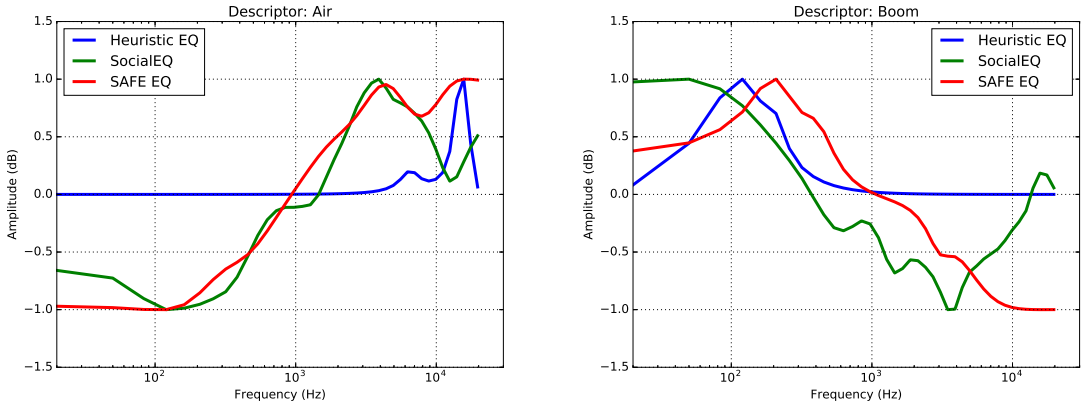
By adjusting the three datasets and extracting the EQ curves corresponding to timbral adjectives, a comparative review between these can be performed. Through this process a number of areas can be explored that would not be possible by examination of a single dataset. Studies of semantic descriptors have attempted to establish a confidence metric (Cartwright and Pardo, 2013; Stables et al., 2016) for terms, in order to find out which adjectives are more agreed upon. This can now be extended by measuring the confidence across all three datasets as well as the extent to which the same adjectives appear in a similar order. This process will explain whether there are terms that inherently display a higher degree of confidence.

In addition, by comparing the different datasets the agreement on the definitions of single terms can be measured. Instead of measuring the confidence of terms within datasets, the mean EQ curve of shared terms can be compared between datasets. Figure 4.8 illustrates the mean EQ curves for the terms *air* and *boom* in the three datasets. In this manner, it can be established if the definition of a term exhibits a universal agreement. Furthermore, the relationship of a specific term to all other terms can be compared between studies. Here, the relationship of a term such as *warm* to all other terms can be examined, and if that is retained across datasets it will suggest that relationships between terms exhibit a universal agreement.

Finally, the overall structure of the three datasets can be compared in order to measure the similarities of distances between terms across datasets. This process will highlight the universal relationship between terms.

Term confidence

Elicited definitions of semantic terms frequently contain diverse spectral representations. For example, the term *anemic* is defined as a cut between 20-110 Hz by Izhaki (2013), while Gibson (2005) defines it as a cut between 40-200 Hz. These differences in definitions can be measured in order to find the extent to which a consensus exists between them. This measurement is performed through the biased sample variance, defined in Equation 4.26.



(a) (b)

Figure 4.8: Mean EQ curves for the terms (a) *air* and (b) *boom*.

$$\hat{\sigma_j}^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{4.26}$$

This calculation of variance can be performed over the frequency bins. The variance is then calculated for each bin (x_i) on term j, and in turn the value is normalised by the number of instances (N) that exist for each term.

Confidence comparison To evaluate whether the ranking of terms, from most to least confident, is retained across the datasets, the common terms between the three datasets are identified and ranked based on their confidence rating. By converting these values to ranks, Spearman correlation (Equation 4.27) can be used to examine the extent to which the order is retained.

$$r_n = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{4.27}$$

where n is the number of observations, and d_i is the difference between two ranks ($rg(X)$ and $rg(Y)$) of each observation, such that:

$$d_i = rg(X_i) - rg(Y_i) \tag{4.28}$$

Definition comparison

The definition comparison aims at uncovering the relationship of terms across studies in three distinct levels in order to fully investigate the EQ curves and the structure of the datasets. These three comparisons are:

- *Coherence between mean definitions*. This is performed in order to find if agreement exists for EQ curves of the same term.
- *Relational distance comparison*. This will establish whether the relationship of a term to all other terms is retained.
- *Structural comparison*. This will provide a measurement of the overall similarity between the structures of the datasets.

Mean definition comparison An essential aspect of this review section is to find the extent to which the terminology definitions are consistent between the different datasets. The extent to which two EQ curves are similar can be measured by using the coherence measurement (Equation 4.29).

$$C_{ab}(f) = \frac{|G_{ab}(f)|^2}{G_a(f)G_b(f)} \tag{4.29}$$

where G_{ab} is the cross-spectral density between equalisation curve a and b, while G_a and G_b the spectral density of the a and b curves respectively.

This method allows us to accurately measure the relationship between shared terms of the three datasets. The metric produces results ranging from -1 to +1, where a strong negative value suggests a high level of dissimilarity between the equalisation curves being measured, and a high positive value suggests a high level of similarity between EQ curves. By also acquiring the p-value from the metric, the significance of the findings can be observed.

Relational distance comparison Following the measurement between the definitions of shared terms in the three datasets, the relationship of each individual term to each other

term can be measured to establish the level of similarity/dissimilarity that exists between them. A symmetric pairwise distance matrix is established with the same terms and dimensions for each dataset. In order to find entry of distance matrix $D(i, j)$ the coherence of term i to term j is measured. However, the coherence score needs to be converted to a distance metric. In order to perform this measurement the coherence values will need to be inverted, with a value of -1 suggesting a perfect positive correlation and a value of +1 a perfect negative correlation, and in turn rescaled to a range of 0 to 2. By comparing the correlation of the individual rows across the three distance matrices, the extent to which the relationships are retained can be established. The resulting distance matrices are displayed in Figure 4.9.

Here, correlation between the term similarities can be assessed by measuring the i^{th} row of the SAFE EQ distance matrix (Figure 4.9a) with the i^{th} row of the SocialEQ (Figure 4.9b) and the Heuristic EQ (Figure 4.9c) matrices. If the measurement produces a positive correlation, it suggests the relationship of term i to the other timbral adjectives is retained across datasets.

Structural comparison By calculating the coherence of each mean definition (Equation 4.29), each term can be measured against all other terms and the resulting distances between representations can be established. The structural similarity between the three datasets can then be measured by acquiring the pairwise distance matrices between the common terms, projecting them to a vector form, and measuring the correlation of the resulting vectors. This is known as the Mantel test (Mantel, 1967), as defined by Legendre and Legendre (2012) in Equation 4.30.

$$r_m = \frac{1}{d-1} \sum_{i=1}^{n-1} \sum j = i + 1^n stand(D_X)_{ij} stand(D_Y)_{ij} \tag{4.30}$$

where d is defined as:

$$d = \frac{n(n-1)}{2} \tag{4.31}$$

and n the number of rows or columns in each of the distance matrices (D_X and D_Y). Finally, the $stand$ function ensures the the distance matrices contain standardized distances. By performing the Mantel test it is possible to assess the correlation of the distances be-

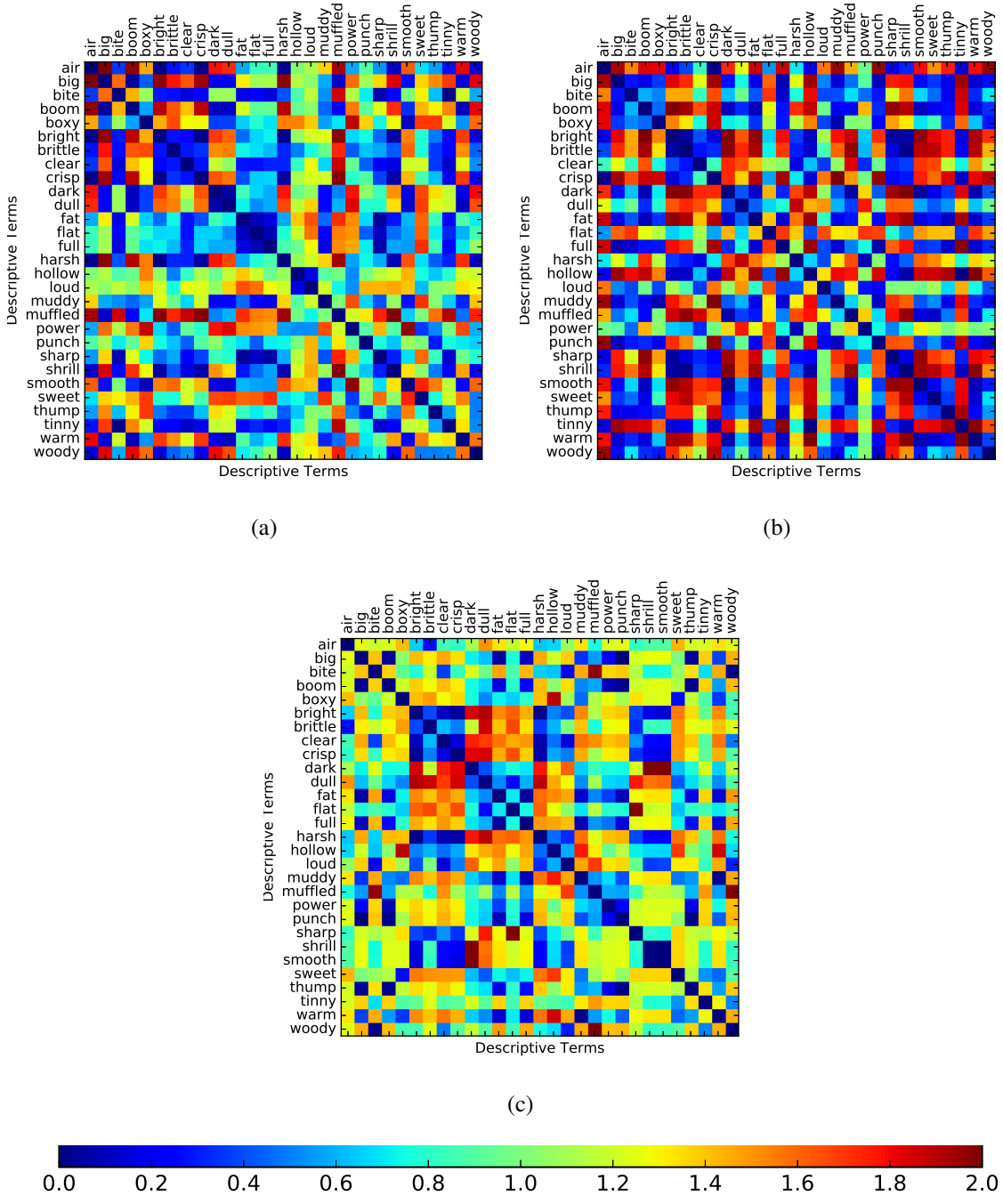


Figure 4.9: Pairwise distance matrices for the (a) SAFE EQ, (b) SocialEQ and (c) Heuristic EQ datasets across shared terms.

tween the three datasets, identifying whether or not these share a common structure.

4.4 Discussion

4.4.1 Confidence comparison

Terms appearing with multiple definitions comprise 47.3% of the Heuristic EQ dataset. The existence of the multiple definitions within that list provides the opportunity of assessing the agreement between entries by measuring the biased sample variance (Equation 4.26). The terms in the Heuristic EQ dataset displaying the lowest variance scores are found to be *shrill* (0.002121), *brittle* (0.002158) and *solid* (0.005162), while the terms displaying the highest variance are *sweet* (0.223748), *crunch* (0.210761) and *horn* (0.149815). The ten most confident and ten least confident terms in the heuristic dataset are displayed in Table 4.3.

Lowest Variance Terms	**Variance Scores**	**Highest Variance Terms**	**Variance Score**
shrill	0.002121	*sweet*	0.223748
brittle	0.002158	*crunch*	0.210761
solid	0.005162	*horn*	0.149815
punch	0.008245	*distant*	0.144803
subsonic	0.008370	*zing*	0.125148
thump	0.012541	*honk*	0.122273
loud	0.013499	*boxy*	0.114336
crisp	0.013636	*nasal*	0.107126
air	0.013668	*close*	0.102161
body	0.016736	*tinny*	0.101004

Table 4.3: A list of the ten most confident, and ten least confident terms in the Heuristic EQ dataset.

Similarly, the crowd-sourced datasets also feature a number of terms, which have multiple definitions in the frequency domain. These can be measured in order to find the extent to which agreement exists between their definitions. Tables 4.4 and 4.5 show the ten most confident and ten least confident terms in the SocialEQ and SAFE EQ datasets respectively.

By implementing the biased sample variance metric, the variance ratings of the terms can be observed. The terms in the SocialEQ dataset exhibiting the lowest variance are

warm (0.0134), *tinny* (0.0193) and *low* (0.0197), while the terms displaying the highest variance are *wild* (0.468), *morning* (0.443) and *dull* (0.415).

Lowest Variance Terms	**Variance Score**	**Highest Variance Terms**	**Variance Score**
warm	0.013370	*wild*	0.467795
tinny	0.019348	*morning*	0.442651
low	0.019749	*dull*	0.415232
dry	0.020644	*welcome*	0.375840
metallic	0.022252	*graceful*	0.357994
loud	0.026638	*love*	0.346433
soft	0.030038	*rock*	0.343432
cold	0.030104	*tense*	0.334179
bright	0.033519	*wet*	0.333322
quiet	0.034225	*hello*	0.310092

Table 4.4: A list of the ten most confident, and ten least confident terms in the SocialEQ dataset.

The ratings for the biased sample variance are illustrated in Table 4.5. The terms that appear most often in the SAFE EQ dataset, *bright* and *warm* display the least variance, 0.0368 and 0.0441 respectively. On the other hand, the terms that exhibit the highest variance are *smooth* (15.1649), *fat* (12.5959) and *quiet* (9.3064).

Lowest Variance Terms	**Variance Score**	**Highest Variance Terms**	**Variance Score**
bright	0.036853	*smooth*	15.16492
warm	0.044140	*fat*	12.59597
thump	0.576402	*quiet*	9.306440
present	0.787746	*hollow*	8.613943
fizzy	0.926072	*tinny*	8.152962
clean	0.932241	*dark*	7.790121
tight	0.988496	*deep*	7.706840
twangy	1.200997	*harsh*	7.050298
air	1.301667	*dull*	7.047658
tight	1.338147	*sharp*	6.642881

Table 4.5: A list of the ten most confident, and ten least confident terms in the SAFE EQ dataset.

As the parameter values differ between the three datasets, it is not possible to perform a direct comparison on the confidence values between terms. For this reason, the common terms between all datasets will be identified and will be assigned ranks as to their place in

the confidence table. By filtering out the terms that are not shared between datasets, and converting the confidence values into ranks, Spearman correlation (Equation 4.27) can be used in order to uncover the extent to which agreement is shared between all three studies.

For the Spearman correlation score between the two crowd-sourced datasets, SocialEQ and SAFE EQ, the comparison scores a value of -0.05913 ($p > 0.05$). For the comparisons between the crowd-sourced data and the heuristic definitions, SocialEQ achieves a score of 0.27272 ($p > 0.05$) while the SAFE EQ displays a negative correlation of -0.41958 ($p > 0.05$). Considering all respective p-values ($p > 0.05$), no significant correlation exists regarding the confidence of terms. This result further suggests that there may not be terms which display a higher agreement score inherently. For instance that a consensus exists regarding the definition of the term *bright*, while the definition of *warm* might be more contested. This would further suggest that the agreement scores on the datasets are circumstantial.

4.4.2 Definition comparison

Mean definition comparison

The similarity between the overall definition of terms can be found by using the coherence metric (Equation 4.29). Through this measurement it can be deduced whether the overall representations of individual terms are consistent across datasets, for example if the mean EQ curve for *warm* in the SAFE EQ dataset correlates with the mean EQ curves of *warm* in the SocialEQ and Heuristic EQ datasets.

Table 4.6 illustrates the correlation scores of the common terms between studies, where it is found that the SocialEQ contains 14 terms that display a significant correlation with Heuristic EQ ($p < 0.05$). Similarly, the SAFE EQ definitions display a significant correlation with Heuristic EQ in 16 cases ($p < 0.05$). In more detail, it is found that for the SocialEQ 62% of the definitions display a positive correlation with Heuristic EQ terms, while 82.7% of the SAFE EQ examples display a positive correlation with the Heuristic EQ. Furthermore, SAFE EQ and SocialEQ display a significant correlation of term definitions for 17 timbral adjectives ($p < 0.05$).

By summing the correlation scores, the overall mean correlation can be found. Through this process it is found that the terms in the Heuristic EQ dataset display a score of +0.2202

Term	Heuristic-Social	Heuristic-SAFE	SAFE-Social	Overall	Variance
air	+0.224	+0.467**	+0.925***	+0.538	0.084
big	+0.871***	+0.71***	+0.868***	+0.816	0.005
bite	+0.132	+0.8***	-0.284	+0.216	0.199
boom	+0.724***	+0.69***	+0.516***	+0.643	0.008
boxy	+0.237	+0.546***	+0.583***	+0.455	0.024
bright	+0.376*	+0.59***	+0.759***	+0.574	0.024
brittle	-0.126	+0.087	+0.487**	+0.149	0.064
clear	-0.126	+0.526***	+0.63***	+0.343	0.111
crisp	+0.66***	+0.423**	+0.914***	+0.666	0.04
dark	+0.198	+0.544***	+0.02	+0.253	0.047
dull	-0.116	+0.747***	-0.412**	+0.072	0.241
fat	+0.69***	-0.277	-0.755***	-0.113	0.36
flat	-0.311	+0.245	-0.222	-0.095	0.059
full	+0.804***	-0.223	-0.283	+0.099	0.248
harsh	-0.06	+0.707***	+0.099	+0.248	0.109
hollow	+0.073	+0.634***	-0.285	+0.14	0.143
loud	-0.238	+0.14	-0.212	-0.103	0.029
muddy	+0.59***	+0.392*	+0.145	+0.375	0.033
muffled	+0.584***	+0.676***	+0.64***	+0.633	0.001
power	-0.016	+0.109	-0.216	-0.04	0.017
punch	+0.753***	+0.208	-0.029	+0.31	0.107
sharp	-0.05	-0.134	+0.836***	+0.217	0.192
shrill	+0.19	+0.291	+0.737***	+0.409	0.054
smooth	-0.103	-0.428**	+0.024	-0.168	0.036
sweet	-0.507***	-0.795***	+0.474**	-0.276	0.295
thump	+0.714***	+0.148	-0.187	+0.224	0.137
tinny	+0.427**	+0.583***	+0.761***	+0.59	0.018
warm	+0.528***	+0.91***	+0.474**	+0.637	0.037
woody	-0.747***	+0.302	-0.606***	-0.350	0.215

Table 4.6: Coherence comparison of the mean EQ curves between datasets with p-value annotations (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

with the Social EQ dataset and a score of +0.3316 with the SAFE EQ dataset. In addition, it is found that the SAFE EQ and Social EQ datasets achieve a mean correlation score of +0.2208. Finally, the mean overall score of all terms across all three datasets is found to be +0.2575. This suggests that there exists an overall agreement between term definitions across studies.

Relational distance comparison

By computing the pairwise distance matrices (Figure 4.9), the relationship of each term to all the other terms in the dataset can be illustrated. In this manner, it can be shown whether not only the definition of a term is retained across the different datasets, but also if its relationship to all other terms is preserved. Table 4.7 presents the resulting correlation scores across the shared terms, where it can be shown that the comparison between the SocialEQ and Heuristic EQ datasets score the most significant results, with a total of 24 terms display a significant score ($p < 0.05$). Similarly the SAFE EQ dataset displays significant results for a total of 19 terms with the Heuristic EQ dataset ($p < 0.05$). Finally, SAFE EQ and SocialEQ display significant results for 21 terms ($p < 0.05$).

In more detail, the comparison between the relationships of individual terms is illustrated in Figure 4.10. Figure 4.10a illustrates the correlation of the term *dull* with all other terms across datasets, with the overall correlation value achieving the highest score across terms (+0.67558). On the other hand, Figure 4.10b, displays the relationship of the term *smooth*, which scores the lowest overall correlation score (-0.15195).

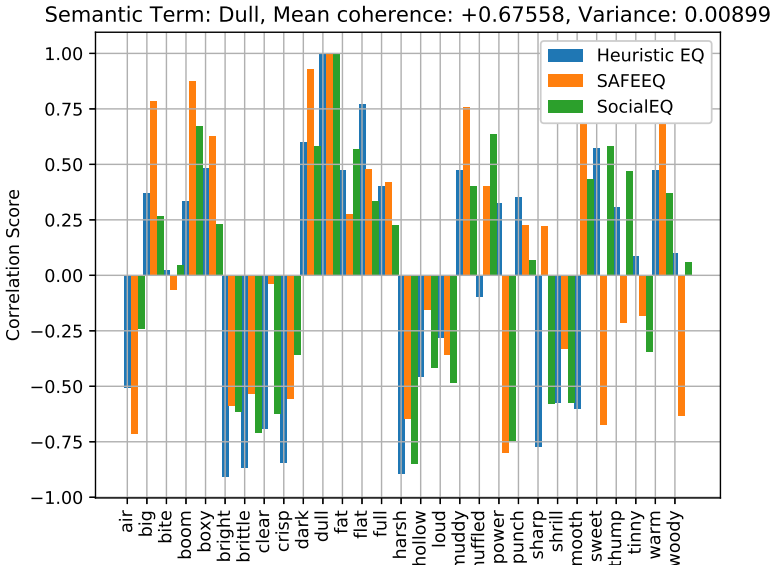
Structural comparison

By converting the coherence measurements (Figure 4.9) to distance values and implementing hierarchical clustering, the structure of the shared terms across datasets can be illustrated in Figure 4.11.

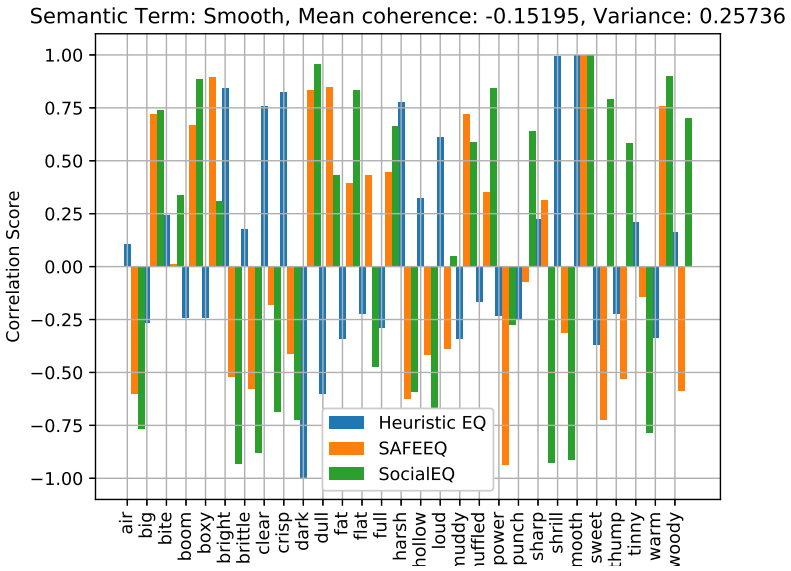
In order to uncover the similarity in the structure of the datasets and the resulting hierarchies, the Mantel test (Equation 4.30) will be implemented to identify the degree to which the relationships are retained. By first measuring the Mantel coefficient between the SAFE EQ and SocialEQ datasets, a score of +0.2897 ($p = 0.002$) is produced, suggesting that a weak positive correlation is present between the two crowd-sourced systems.

Term	Heuristic-Social	Heuristic-SAFE	SAFE-Social	Overall	Variance
air	+0.684***	+0.445*	+0.615***	+0.581	0.01
big	+0.628***	+0.537**	+0.617***	+0.594	0.001
bite	-0.211	+0.399*	-0.383*	-0.065	0.112
boom	+0.585***	+0.533**	+0.704***	+0.607	0.005
boxy	+0.578**	+0.511**	+0.402*	+0.497	0.005
bright	+0.700***	+0.563**	+0.682***	+0.649	0.003
brittle	+0.682***	+0.47**	+0.71***	+0.621	0.011
clear	+0.671***	+0.45*	+0.604***	+0.575	0.008
crisp	+0.61***	+0.558**	+0.606***	+0.591	0.001
dark	+0.621***	+0.496**	+0.61***	+0.576	0.003
dull	+0.772***	+0.546**	+0.708***	+0.675	0.008
fat	+0.713***	-0.168	-0.239	+0.101	0.187
flat	-0.153	+0.046	+0.258	+0.05	0.028
full	+0.656***	-0.002	-0.144	+0.17	0.121
harsh	+0.730***	+0.571**	+0.671***	+0.657	0.004
hollow	+0.646***	+0.435*	+0.232	+0.438	0.028
loud	+0.145	+0.105	+0.123	+0.124	0.001
muddy	+0.689***	+0.45*	+0.384*	+0.508	0.017
muffled	+0.385*	+0.41*	+0.672***	+0.489	0.016
power	-0.221	-0.187	+0.569**	+0.053	0.133
punch	+0.552**	+0.207	-0.053	+0.235	0.061
sharp	+0.637***	+0.165	+0.358	+0.387	0.037
shrill	+0.614***	+0.485**	+0.62***	+0.573	0.003
smooth	-0.595***	-0.418*	+0.558**	-0.151	0.257
sweet	+0.588***	-0.312	-0.277	-0.001	0.173
thump	+0.602***	-0.16	-0.436*	+0.001	0.193
tinny	+0.372*	+0.347	+0.574**	+0.431	0.01
warm	+0.699***	+0.657***	+0.612***	+0.656	0.001
woody	-0.159	+0.387*	-0.51**	-0.094	0.136

Table 4.7: Relational distance comparison of the three datasets with p-value annotations (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).



(a)



(b)

Figure 4.10: Correlation comparison of relational distances between datasets for the term (a) *dull*, displaying the highest overall correlation score, and (b) *smooth*, displaying the lowest overall correlation score.

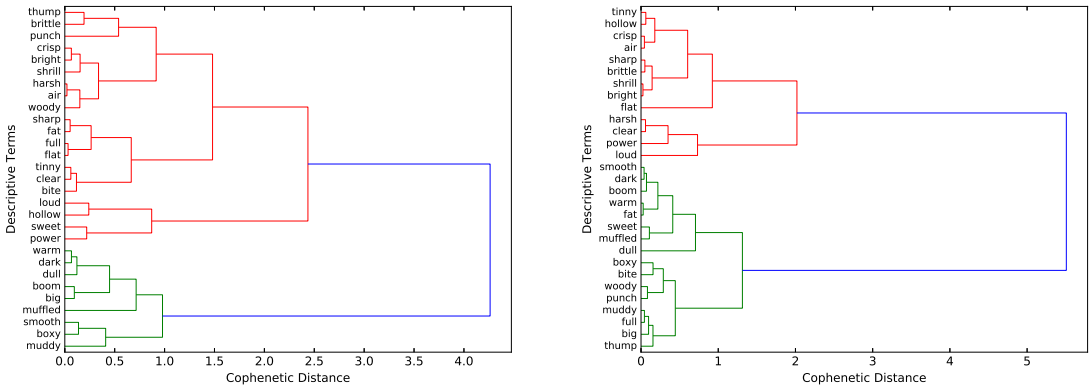


Figure 4.11: Hierarchical clustering of the common terms for the (a) SAFE EQ, (b) SocialEQ and (c) Heuristic dataset.

The measurement between the Heuristic EQ values and the SocialEQ terms achieves a score of +0.4299 ($p = 0.001$), while when comparing the SAFE EQ to the Heuristic EQ structures the score drops to +0.2216 ($p = 0.002$). These results suggest that both datasets are capable of retaining the Heuristic EQ relationships, with SocialEQ displaying a moderate to strong positive correlation, while the SAFE EQ a weak positive correlation.

4.4.3 Global definitions of semantic terms

By normalising the range across the Heuristic EQ, SAFE EQ and SocialEQ datasets, it is possible to construct the mean EQ curves of all shared terms. Through these global definitions, further analysis can be undertaken in the relationships between the gathered terms, and the basis for a semantic thesaurus can be established, the need for which was expressed by Toulson (2006).

Global definition

Firstly, by measuring the pairwise coherence across all global definitions a distance matrix can be constructed (Figure 4.12), where the initial groupings between terms can be illustrated. For instance, from the figure it can be shown that the terms *bright, brittle, clear* and *crisp* display high positive correlation scores, while at the same time this group of terms exhibits high negative correlation scores with the terms *dark, dull* and *fat*.

Table 4.8 displays the significant synonyms, exhibiting a positive correlation with a highly significant p-value (< 0.001), and the significant antonyms, which display a negative correlation with a highly significant p-value (< 0.001). The term synonymity showcases expected results, when considering the placement of the terms in the hierarchy of the individual datasets. For example, the term *air* can be considered a synonym for terms such as *bright* and *harsh*. On the other hand, it is also found that opposing term meanings also display a highly significant opposing relationship, as is the case with *bright* and *dark*, as well as *clear* and *muffled*, or *smooth* and *harsh*. This result suggests that linguistic antonyms are also mirrored in the semantic terminology used by music producers and sound engineers. This in turn would suggest that the understanding of linguistic term definition can potentially be used for simplifying EQ operations. However, apart from the mirroring of the linguistic associations, the resulting table displays opposing relationships of terms that are

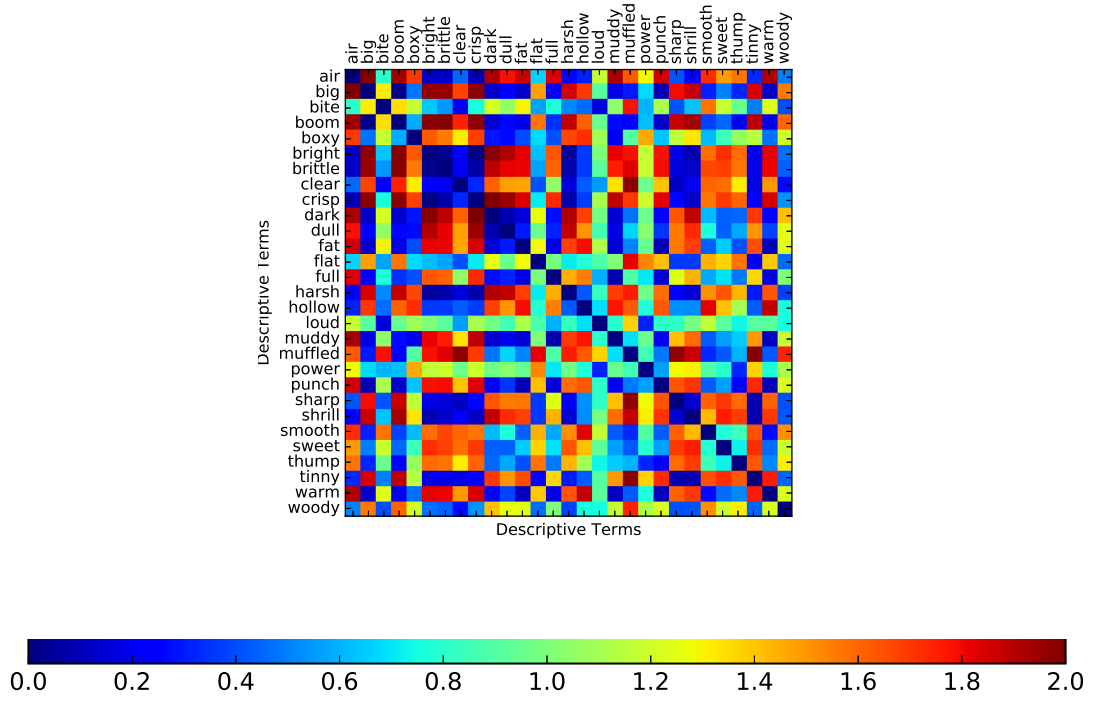


Figure 4.12: Coherence measurements for the global definitions derived from the SAFE EQ, Social EQ and Heuristic EQ datasets.

known to yield opposing timbral characteristics in music production, for example *bright* and *warm*.

Term	**Significant Synonyms**	**Significant Antonyms**
air	bright, brittle, clear, crisp, harsh, hollow, sharp, shrill, tinny, woody	big, boom, boxy, dark, dull, fat, full, muddy, muffled, punch, smooth, thump, warm
big	boom, boxy, dark, dull, fat, full, muddy, muffled, punch, smooth, sweet, thump, warm	air, bright, brittle, clear, crisp, harsh, hollow, sharp, shrill, tinny, woody
bite	clear, hollow, loud, sharp, tinny, woody	muffled, smooth

boom	big, dark, dull, fat, full, muddy, muffled, punch, smooth, sweet, thump, warm	air, bright, brittle, clear, crisp, flat, harsh, hollow, sharp, shrill, tinny, woody
boxy	big, dark, dull, fat, full, muddy, warm	air, bright, brittle, crisp, harsh, hollow
bright	air, brittle, clear, crisp, harsh, hollow, sharp, shrill, tinny, woody	big, boom, boxy, dark, dull, fat, full, muddy, muffled, punch, smooth, sweet, thump, warm
brittle	air, bright, clear, crisp, harsh, hollow, sharp, shrill, tinny, woody	big, boom, boxy, dark, dull, fat, full, muddy, muffled, punch, smooth, sweet, thump, warm
clear	air, bite, bright, brittle, crisp, flat, harsh, hollow, sharp, shrill, tinny, woody	big, boom, dark, muffled, smooth, sweet
crisp	air, bright, brittle, clear, harsh, hollow, sharp, shrill, tinny	big, boom, boxy, dark, dull, fat, full, muddy, muffled, punch, smooth, sweet, thump, warm
dark	big, boom, boxy, dull, fat, full, muddy, muffled, punch, sweet, thump, warm	air, bright, brittle, clear, crisp, harsh, hollow, sharp, shrill, tinny
dull	big, boom, boxy, dark, fat, full, muddy, punch, sweet, warm	air, bright, brittle, crisp, harsh, sharp, shrill
fat	big, boom, boxy, dark, dull, full, muddy, punch, smooth, thump, warm	air, bright, brittle, crisp, harsh, hollow, sharp, shrill, tinny
flat	clear, sharp, shrill, tinny, woody	boom, muffled, thump
full	big, boom, boxy, dark, dull, fat, muddy, punch, thump, warm	air, bright, brittle, crisp, hollow
harsh	air, bright, brittle, clear, crisp, hollow, sharp, shrill, tinny, woody	big, boom, boxy, dark, dull, fat, muddy, muffled, punch, sweet, warm

hollow	air, bite, bright, brittle, clear, crisp, harsh, hollow, tinny	big, boom, boxy, dark, fat, full, muddy, muffled, punch, smooth, warm
loud	bite, power	—
muddy	big, boom, boxy, dark, dull, fat, full, punch, smooth, warm	air, bright, brittle, crisp, harsh, hollow, shrill
muffled	big, boom, dark, punch, smooth, sweet, warm	air, bite, bright, brittle, clear, crisp, flat, harsh, hollow, sharp, shrill, tinny, woody
power	loud, thump	—
punch	big, boom, dark, dull, fat, full, muddy, muffled, smooth, thump, warm	air, bright, brittle, crisp, harsh, hollow, sharp, shrill, tinny
sharp	air, bite, bright, brittle, clear, crisp, flat, harsh, shrill, tinny, woody	big, boom, dark, dull, fat, muffled, punch, smooth, sweet, thump, warm
shrill	air, bright, brittle, clear, crisp, flat, harsh, sharp, shrill, tinny, woody	big, boom, dark, dull, fat, muddy, muffled, punch, sweet, thump, warm
smooth	big, boom, fat, muddy, muffled, punch, smooth, warm	air, bite, bright, brittle, clear, crisp, hollow, sharp, tinny
sweet	big, boom, dark, dull, muffled, warm	bright, brittle, clear, crisp, harsh, sharp, shrill, tinny
thump	big, boom, dark, fat, full, power, punch	air, bright, brittle, crisp, flat, sharp, shrill
tinny	air, bite, bright, brittle, clear, crisp, flat, harsh, hollow, sharp, shrill, woody	big, boom, dark, fat, muffled, punch, smooth, sweet, thump, warm
warm	big, boom, boxy, dark, dull, fat, full, muddy, muffled, punch, smooth, sweet	air, bright, brittle, crisp, harsh, hollow, sharp, shrill, tinny

woody	air, bite, bright, brittle, clear, flat, harsh, sharp, shrill, tinny	big, boom, muffled

Table 4.8: Thesaurus containing significant synonyms and antonyms for the global definitions of equalisation settings.

Relationship to frequency ranges

The proposed problem of "multiple languages", introduced by Izhaki (2013), can also be addressed by considering the relationship of the global definitions of timbral adjectives to the definitions of the ranges of the frequency spectrum. De Man (2017) has also compiled a list of frequency range definitions across the existing literature. Here, it is possible to illustrate the proposed ranges which define each sub-division of the frequency domain, in order to create a semantic dictionary between timbral adjectives and frequency ranges. These are illustrated in Figure 4.13, for the low-end, low-mid, high-mid and high-end.

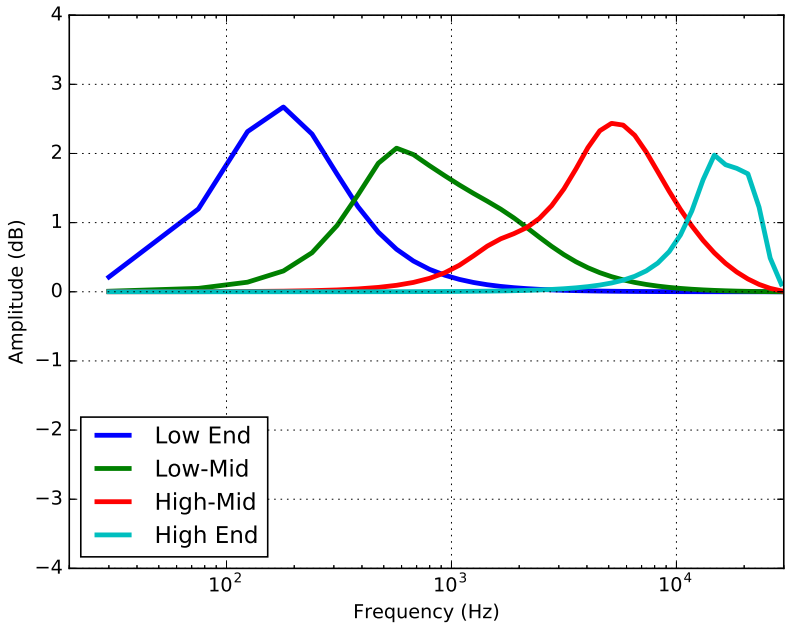


Figure 4.13: Separation of the frequency spectrum into individual ranges per the guidelines of the Heuristic dataset, as provided by De Man (2017).

By measuring the correlation between each frequency range and each semantic term, the timbral adjectives can be categorised to the range(s) in which they operate. Table 4.9 displays the relationship of each global EQ curve to each of the four heuristic frequency ranges. High positive and negative values suggest that the specific term is active in these areas of the frequency range, either in boosting or cutting the frequency components. For instance, a boost in the low-end can be performed in order to add more *boom* ($+0.92828$) to the input sound, or to make something *warm* ($+0.70130$). Similarly, a cut in the same range can be performed to increase the *air* (-0.75150) or to make something more *bright* (-0.76731).

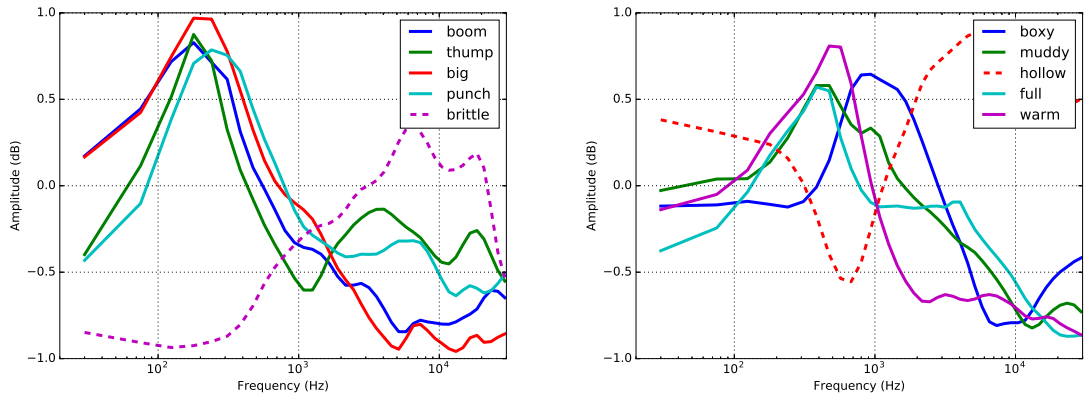
By finding the absolute values of Table 4.9, the semantic terms mostly associated with each range can be visualised. In this context, the five most prominent semantic terms are illustrated in Figures 4.14, displaying both amplifications and attenuations of the frequency ranges. In Figure 4.14a it is shown that the terms displaying the highest positive correlation with the low-end range are *boom, thump, big* and *punch*, suggesting that these timbral adjectives can be implemented by increasing the gain of the low-end. On the other hand, the semantic term displaying a high negative correlation in the same context is *brittle*, a finding that suggests certain terms can be achieved through the attenuation of the low-end frequencies.

On the other hand, for the low-mid frequency range (Figure 4.14b), it is found that the associated terms are *muddy, boxy, warm* and *full*, all of which are achieved through amplification of the frequency components. However, the term *hollow* exhibits a high negative correlation with the frequency region. Similarly, this is the case with the high-mid frequency range (Figure 4.14c), where the terms *bite, harsh, clear, brittle* all display a high positive correlation, while the term *muffled* scores a high negative correlation. Finally, the timbral adjectives correlating with the high-end range (Figure 4.14d) all display a high negative correlation with the frequency region, and these are the terms *full, dull, muddy, loud* and *boxy*.

Table 4.10 shows the relationship of frequency ranges to terms. In this manner, a semantic dictionary is constructed which can connect timbral adjectives with frequency ranges. This offers a quick translation between two different sets of "languages". For example if a sound engineer suggests making something sound *brighter* then this can be done by amplifying the high-mid or high-end frequency ranges.

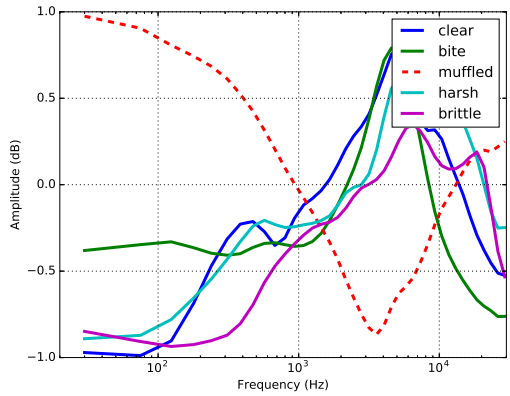
Term	Low End	Low-Mid	High-Mid	High End
air	-0.751***	-0.586***	+0.563***	+0.567***
big	+0.892***	+0.463**	-0.608***	-0.508***
bite	-0.225	-0.069	+0.894***	-0.414**
boom	+0.928***	+0.314*	-0.638***	-0.436**
boxy	+0.197	+0.866***	-0.443**	-0.572***
bright	-0.767***	-0.416**	+0.721***	+0.495**
brittle	-0.786***	-0.374*	+0.767***	+0.464**
clear	-0.558***	-0.075	+0.917***	-0.034
crisp	-0.755***	-0.485**	+0.639***	+0.57***
dark	+0.706***	+0.566***	-0.591***	-0.527***
dull	+0.572***	+0.531***	-0.435**	-0.732***
fat	+0.758***	+0.662***	-0.583***	-0.405**
flat	-0.591***	+0.318*	+0.457**	-0.358*
full	+0.587***	+0.709***	-0.119	-0.734***
harsh	-0.637***	-0.348*	+0.799***	+0.402*
hollow	-0.401*	-0.790***	+0.731***	+0.207
loud	+0.125	-0.029	+0.671***	-0.597***
muddy	+0.56***	+0.829***	-0.396*	-0.714***
muffled	+0.669***	+0.037	-0.838***	+0.077
power	+0.544***	-0.339*	+0.301	-0.235
punch	+0.882***	+0.474**	-0.387*	-0.519***
sharp	-0.757***	+0.004	+0.76***	+0.127
shrill	-0.783***	-0.125	+0.677***	+0.225
smooth	+0.513***	+0.491**	-0.621***	-0.195
sweet	+0.401*	+0.156	-0.402*	-0.354*
thump	+0.898***	-0.095	-0.181	-0.237
tinny	-0.782***	-0.085	+0.699***	+0.077
warm	+0.701***	+0.684***	-0.517***	-0.481**
woody	-0.482**	+0.101	+0.647***	-0.127

Table 4.9: Correlation between semantic terms and frequency ranges with p-value annotations (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

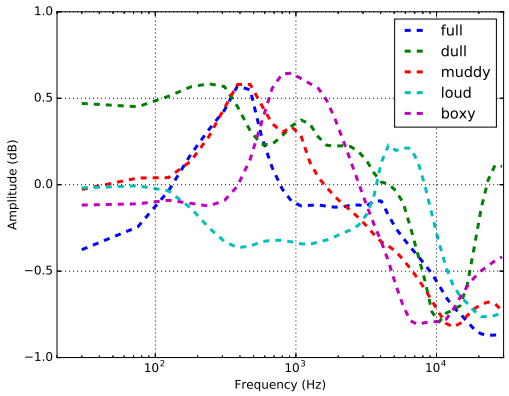


(a)

(b)



(c)



(d)

Figure 4.14: Frequency distributions of prevalent terms for the low-end (a), low-mid (b), high-mid (c) and high-end (d) ranges. Terms that display a positive correlation with the frequency range are depicted as solid lines, while terms that display negative correlation with the frequency ranges are depicted as dashed lines

Frequency Range	Significant Positive Correlating Terms	Significant Negative Correlating Terms
Low End	big, boom, boxy, dark, dull, fat, full, loud, muddy, muffled, power, punch, smooth, sweet, thump, warm	air, bite, bright, brittle, clear, crisp, flat, harsh, hollow, sharp, shrill, tinny, woody
Low-Mid	big, boom, boxy, dark, dull, fat, flat, full, muddy, muffled, punch, sharp, smooth, sweet, war, woody	air, bite, bright, brittle, clear, crisp, harsh, hollow, loud, power, shrill, thump, tinny
High-Mid	air, bite, bright, brittle, clear, crisp, flat, harsh, hollow, loud, power, sharp, shrill, tinny, woody	big, boom, boxy, dark, dull, fat, full, muddy, muffled, punch, smooth, sweet, thump, warm
High End	air, bright, brittle, crisp, harsh, hollow, muffled, sharp, shrill, tinny	big, bite, boom, boxy, clear, dark, dull, fat, flat, full, loud, muddy, power, punch, smooth, sweet, thump, warm, woody

Table 4.10: Categorisation of terms by frequency range. A term is considered significant for a range if its correlation score achieves a p-value of $p < 0.05$. Then depending on the nature of the correlation it is assigned to either the positively or negatively correlating terms.

4.5 Conclusion

This chapter presents an analysis of three different datasets containing semantic terms associated with EQ parameters. Connections between term definitions, relationships between terms and structural similarities have been identified, suggesting that global representations of timbral adjectives exist between studies. Finally, the chapter has demonstrated the methodology for defining a *thesaurus* of semantic terms and a dictionary mapping timbral adjectives to frequency ranges.

4.5.1 Term confidence

The confidence of term definitions was measured in all three datasets through the biased sample variance (Equation 4.26) and the resulting term vectors (Tables 4.3, 4.4 and 4.5) were compared through the Spearman correlation. The results of this process suggest that the confidence between terms is dependent on the specific dataset, and is not displayed across datasets. All comparisons of the Spearman correlation between confidence vectors display a p-value > 0.05, assessing the results as non-significant. This suggests that there are no terms that are inherently more agreed upon than others.

4.5.2 Definition comparison

In the previous sections the commonality of terms and the confidence of individual terms has been questioned. This section focuses on the comparison of individual definitions, the relationship between definitions, and the structural information of the datasets. In the context of these attributes there appears to be an overall agreement.

Mean definition comparison

By investigating the coherence measurements for the shared terms, the extent to which agreement exists between definitions is found. By summing the coherence measurement for the common terms, the overall coherence of each dataset is established, with SAFE EQ achieving a score of +0.3316 and SocialEQ a score of +0.2202 respectively when compared to Heuristic EQ terms. In addition, it is found that the SAFE EQ and Social EQ terms achieve an overall correlation of +0.2208. As all comparison combinations between datasets display a positive mean coherence score and a mean overall correlation of +0.2575, it can be said that an agreement exists regarding the individual spectral definitions of semantic terms.

Relational definition comparison

By examining the relationship of each individual term to all other terms across datasets, the preservation of relations between timbral adjectives has been established. The relationships of terms between the SocialEQ and Heuristic EQ datasets achieve a mean coherence score

of +0.4562, while SAFE EQ and Heuristic EQ achieve a score of +0.2943. Finally the relationships between SAFE EQ and SocialEQ datasets achieve a score of +0.3397. The overall mean coherence score in this case is +0.3634, which suggests that relationship between terms are preserved across datasets.

Structural comparison

The structural similarity across datasets was measured through the Mantel test. Here, the relationships between the shared terms of the three studies is assessed. Based on the respective distance matrices, it is found that SocialEQ achieves a score of +0.4299 to the Heuristic EQ dataset, while SAFE EQ achieves a score of +0.2216 when compared to the Heuristic EQ structure. Finally, the structural assessment between SocialEQ and SAFE EQ datasets achieves a score of +0.2897. The mean Mantel score is found to be +0.3137, which suggests that the underlying structure between term distributions is shared between studies.

4.5.3 Dictionaries of terms

This chapter has performed a comparative review on the existing datasets regarding semantic terminology used in equalisation. In that regard, it has established the similarities and dissimilarities between term entries. By extracting the definitions of all shared terms across the datasets, the global definitions of all terms were sourced. In addition a similarity table was composed, which will provide guidance as to the semantic synonyms and antonyms of all global definitions, and can be used as the basis for a universal terminology in sound engineering (Toulson, 2006), in the context of equalisation. Furthermore, an additional measurement was performed to establish the relationship of global definitions to the different ranges of the frequency spectrum, providing a solution to the problem of "multiple languages" used in music production practices (Izhaki, 2013).

4.5.4 Multiple representations of terms

Whilst this chapter has provided a comparative review between the definitions of semantic terminology across datasets, it has not analysed the existence of multiple definitions for a

single term. Any semantic term can be achieved through a number of different processes. This was also shown in Chapter 3, where different processing chains were implemented in order to achieve the same timbral adjective. The existence of multiple definitions is also highlighted in the existing literature definitions gathered by De Man (2017), and therefore instead of finding the overall distribution, each definition can be treated as a different sub-representation of the overall term. The following chapter will perform an analysis on these sub-representations for the terms *bright* and *warm*.

4.5.5 Limitations

While this study has effectively performed a comparative review of three prominent datasets containing semantic terms, there still exist certain limitations that arise from the format of the datasets. All three datasets implement different methods for gathering equalisation parameters, and they also use different equalisation methods and designs to alter the timbre of an input sound. Even though these were brought to a comparable level to perform comparisons, there is inadvertently some loss of information in the process.

Furthermore, even though this chapter has uncovered a similarity with regard to the definitions, relational distances and structural information between the three datasets, it is unclear whether that will still be the case if a new dataset of semantic terms is presented. Since the current pool of datasets is limited, further information will need to be gathered in order to asses whether these different degrees of similarity still hold.

Chapter 5

Descriptor sub-representations

5.1 Introduction

The comparative review performed in Chapter 4 detailed the relationships between terms, resulting in a thesaurus of semantic equalisation terminology. However, it was also found that the confidence of individual terms differs significantly between datasets. If the confidence of a timbral adjective is low, it suggests that different interpretations of EQ parameters can be used in order to achieve that adjective. In that manner, many semantic terms can be achieved through a number of different processes. Some of these were already illustrated in Chapter 3 where it was found that the same term could be achieved through a number of different combinations of audio effects. The same can occur with different EQ parameters being used to achieve the same term and in order to investigate this behaviour individual timbral adjectives need to be studied in isolation. For this reason, this chapter presents a method for partitioning semantic term definitions into sub-representations.

A number of studies in semantic terminology have identified term definitions and synonymity between terms (Seetharaman and Pardo, 2014a; Cartwright and Pardo, 2013), but they do not account for variations of the same term. Considering the disagreement on term definitions between listeners (Darke, 2005), even between individuals who speak the same language (Disley and Howard, 2004), it is essential to perform an analysis on subsets of single terms.

Differing definitions of the same described equalisation curves are also given throughout audio engineering literature (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins

and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005), with sound engineers and music producers offering a wide number of parametrisations for the same term. This point is further reinforced by Katz (2003), where the author suggests that an inverse action can achieve the same timbral characteristic. For instance, in order to make something *brighter* one approach would be to boost the high end of the frequency spectrum, but an alternate method would be to decrease the low end presence. This approach suggests that the disagreement of subjects regarding timbral adjectives is not only caused due to perceptual differences, but also due to using different methodologies to achieve the same timbral characteristic.

For this reason, as an alternative to synonymous categorisation, it is possible to consider that an underlying structure can exist within a single term. In this example, a term such as *dark* would not be considered synonymous to *warmth*, but rather its sub-representation. In the context of equalisation it is hypothesised that it is possible to uncover the underlying relationships between different definitions of the same term, through the EQ curves implemented by the users. In this way, the problem of multiple definitions is addressed, where each sub-representation can be considered a different route in achieving an overarching, parent descriptor.

For this research two terms are considered, *warm* and *bright*. *Warm* and *bright* are frequently investigated in existing research. *Bright* often exhibits a correlation with the spectral centroid (Beauchamp, 1982) and the fundamental frequency (Marozeau et al., 2003; Marozeau and de Cheveigné, 2007), whereas *warm* has produced antithetical results. Ethington and Punch (1994) consider *warm* to be a result of the energy in the first three harmonics of a signal. Alternatively, Zacharakis (2013) shows the term to display correlation with different measurements of the tristimulus, and considers *warm* to be an opposing timbral characteristic to *bright*. Disley and Howard (2004) further suggest that if this opposing relationship between *warm* and *bright* is true, then *warm* should display a correlation with the fall off of the spectral centroid.

In the context of EQ curves, the sound engineering literature (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005) also displays disagreements over the definitions of frequency ranges for achieving the two terms. As was performed in Chapter 4, the heuristic definitions from the literature can be projected as filter frequency responses.

Figure 5.1 displays the *bright* EQ curves generated from the guidelines provided by Katz (2003), Izhaki (2013) and Gibson (2005). Similarly, Figure 5.2 shows the *warm* EQ curves generated from the guidelines provided by Owsinski (2013), Izhaki (2013), Huber and Runstein (2013), Gibson (2005), Cousins and Hepworth-Sawyer (2013) and Katz (2003). In both cases, for the various definitions for *bright* and *warm*, it is not necessary to consider that the individual definitions from the literature are conflicting with each other, but that each different EQ curve offers a different interpretation of the same term. Therefore, the above EQ curves can be considered as simply subsets of the two terms, which can be named sub-representations.

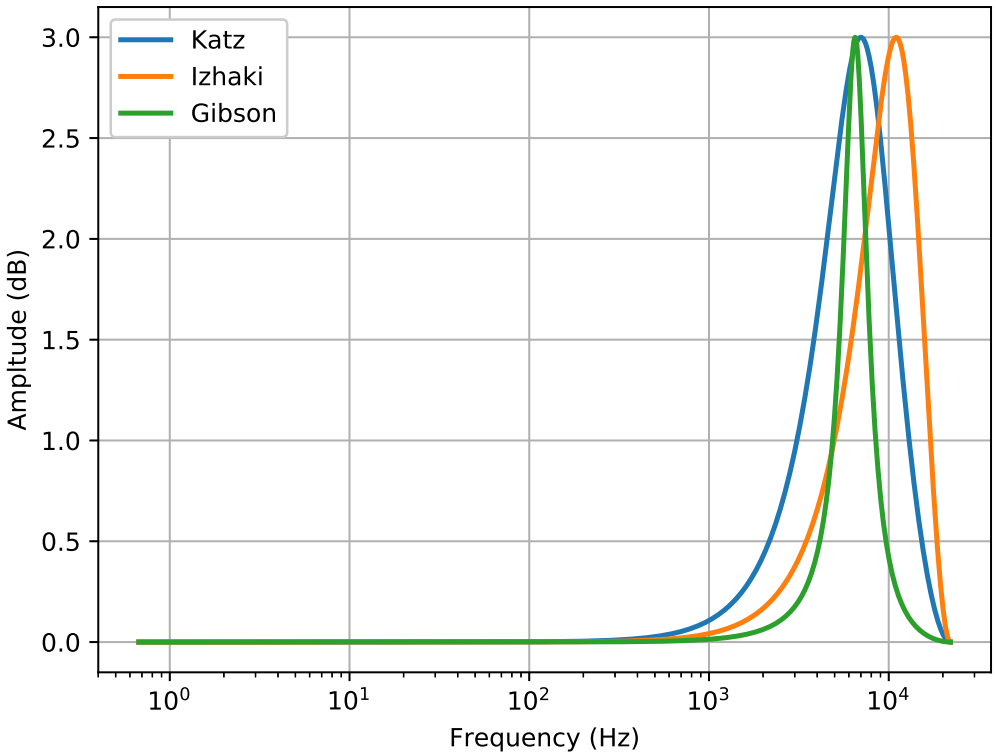


Figure 5.1: EQ curves for *bright* generated from the heuristic definitions gathered by De Man (2017)

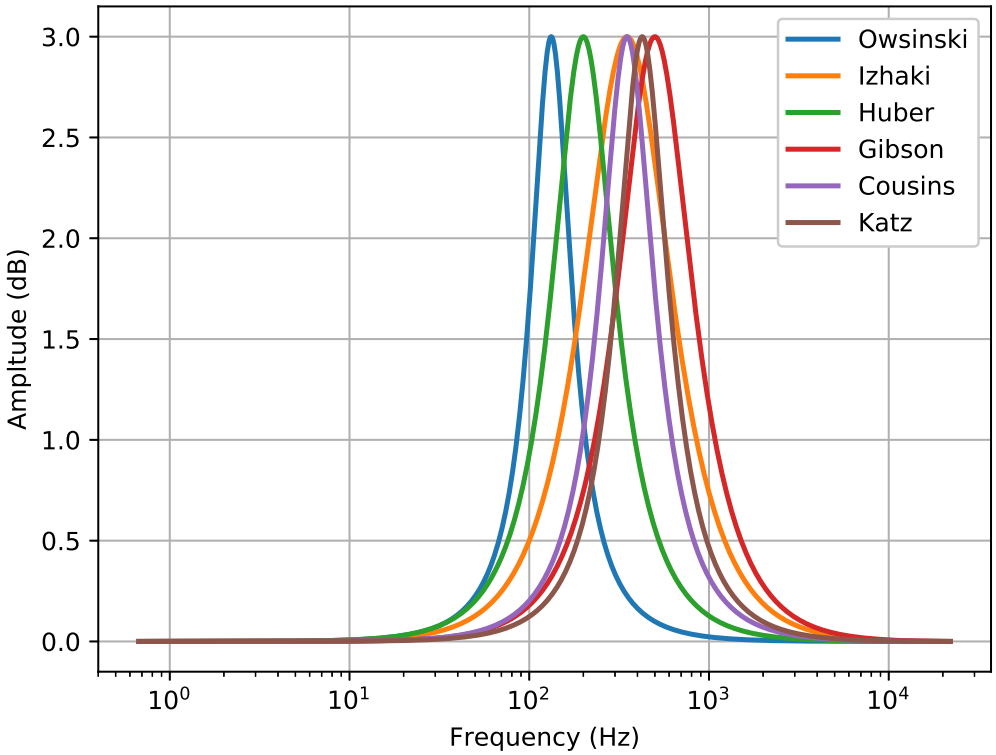


Figure 5.2: EQ curves for *warm* generated from the heuristic definitions gathered by De Man (2017)

5.2 Experiment Design

The dataset for identifying sub-representations of EQ parameters was gathered through the SAFE Project (which was discussed in Sections 2.2.2 and 4.3.1), and two terms chosen for this research were *warm* and *bright*. These were chosen due to their frequent appearance in the existing research and literature, their broad use by producers, engineers and musicians, and finally as the two most frequent terms appearing in the SAFE EQ dataset. The dataset is comprised of EQ parameters, with 13 dimensions in total, as these were illustrated in Table 4.1.

The terms are evaluated in isolation, resulting in two distinct datasets, one for the *bright* entries and another for the *warm* entries, which will be processed and evaluated in the same

manner. The *bright* dataset features 435 instances, while the *warm* dataset features 442 instances. Through this methodology, sub-representations of each descriptor will not be influenced by the existence of a second descriptor in the same dataset.

Semantic equalisation, even though it features certain guidelines for achieving timbral adjectives, is input dependent. As such, in the process of making something *warm* or *bright*, subjects may perform actions that are reactive or adaptive to the audio signal. For this reason, a smoothing process needs to be introduced that will retain all salient information, whilst filtering out the signal-specific attributes. For example, if a user makes a sharp cut in order to reduce some problematic frequencies from the input audio, then that information should not be considered essential in achieving the timbral adjective. In order to implement this processing stage, a model of stacked autoencoders (sAE) is considered, which was introduced in Section 2.1.2.

Once the smoothing stage has been performed, a clustering technique will be implemented in order to partition the dataset into its sub-groups. Despite the advantages that clustering algorithms offer (ease of implementation, unsupervised architecture), they require parameter optimisation, choice for the resulting number of clusters and display difficulties in performing a comparative review between different techniques. All these aspects are discussed in depth in Section 5.3.2. It is considered that the most important factor to be addressed in the use of clustering is the metric implemented to perform the partitioning of the dataset. For this reason a distance metric that is relevant to the type of data in this research is used. By converting the smoothed EQ parameters obtained from the sAE model into EQ curves, the coherence between these can be measured. The method for performing this modified clustering is presented in Section 5.4.2.

Finally, once the appropriate clusters are identified, the resulting sub-representations will be used for a subjective evaluation task. The hypothesis is that if the partitioning process was implemented correctly, then subjects should be able to perceive significant differences between sub-representations of each term. That will suggest that each term features a number of perceptually distinct forms, which can explain the disagreement of subjects regarding the definition of each term.

In order to evaluate the extent to which the system is able to uncover underlying sub-representations, a MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) test is implemented through the Web Audio Evaluation Toolbox (Jillings, Moffat, De Man, Reiss

and Stables, 2016). Here, the participants were asked to rate the acoustic similarity of a reference sound to other versions of the same sound, processed with randomized equalisation curves from the cluster centroids of the prominent sub-representations.

Each slider had a range of 0 to 1, with 0 suggesting no similarity between reference and sample, and 1 suggesting perfect similarity between reference and sample. In total, 20 subjects participated in the listening tests, with varying levels of experience (0-5 years) in professional audio production. Subjects were aged between 18-40, and did not display any hearing impairment.

The existence of sub-representations can therefore be proved perceptually if the subjects are able to distinguish sonic differences between their application on the same sample. For each test partition two electric guitar recordings were presented to the subjects, one performed in a *metal* style and one in a *blues* style.

After being processed through equalisation, the samples were normalised in order to prevent subjects from identifying differences due to the volume inconsistency. The number of samples presented at any one time was dependent on the number of clusters detected for each term. Finally, a similarity matrix of distance vectors is constructed from each reference audio sample.

Figure 5.3 presents the stages for identifying sub-representations. The pre-processing stage introduces the dataset, comprised of 13 parameters, which is passed through the sAE model in order to smooth specificities from the input audio. In order to then partition the dataset, the 13 parameters are converted to EQ curves represented as frequency bins, which are then passed to a clustering algorithm. Through this process, the centroids of the different subsets of the EQ curves are identified, and are used for the subjective evaluation task.

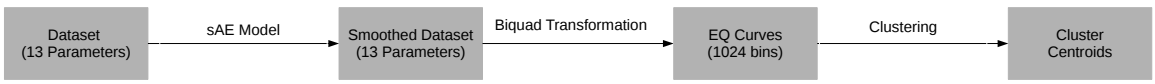


Figure 5.3: Experiment design flowchart for reconfiguring the data and detecting sub-representations. The process involves two transformation tasks (sAE, Biquad transformation) and a clustering step.

5.3 Techniques

5.3.1 Stacked autoencoders

In order to deal with specificities of EQ parameters in the SAFE dataset, a model of sAE is considered, which essentially performs dimensionality reduction and reconstruction on an input dataset.

Hinton and Salakhutdinov (2006) initially proposed a method for reducing dimensionality through the use of ANNs, more specifically autoencoders. An autoencoder, also known as auto-associator or Diabolo network (Bourlard and Kamp, 1988; Hinton and Zemel, 1994; Rumelhart et al., 1988), is a type of ANN with the function of learning a compact representation of a given dataset. This is performed by optimising a matrix of weights and bias units, with such values that a loss function, representing the difference between the input and output (reconstructed input) is minimal. Autoencoders can also be stacked in series, using the output of the prior layer as the input of the next, to build a deep network architecture (sAE). Autoencoders also feature a less tedious process of training in which the individual layers are trained separately and then the whole system is fine-tuned. Due to this lighter process autoencoders have been used as the basis for deep architectures (Bengio et al., 2007; Larochelle et al., 2007).

Each layer of the model is trained individually, minimising the reconstruction error between its input and output stages (Figure 5.4). This method has been used for data compression, and by extension, for dimensionality reduction. A common use of sAE is for improving the classification accuracy of logistic regression (Bengio et al., 2009), but a final reconstruction layer can also be implemented in order to retrieve an approximation of the input.

The model is trained to encode a given input x into a representation $c(x)$, so that the representation can accurately reconstruct the input. The preferred method for optimising a model of sAE is by minimising the negative log-likelihood error criterion (Equation 5.1).

$$RE = -logP(x|c(x)) \tag{5.1}$$

Then the loss function is defined as in Equation 5.2, where f represents the decoder function and $f(c(x))$ the reconstruction process followed by the system, obtained through

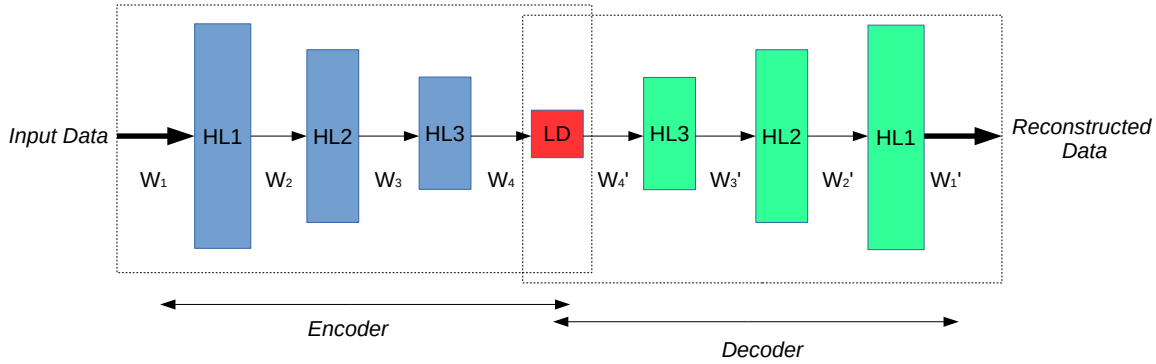


Figure 5.4: Schematic representation of a sAE model. Here, the input data (x) are processed through a series of hidden layers (HL), comprised of weight matrices (W_i), to produce a low-dimensional mapping (LD). The process can then be reversed in the decoder part of the system, by transposing the weights (W_i') of the hidden layers to retrieve the reconstructed dataset ($c(x)$).

a sigmoid function, and i the layer of the sAE being trained.

$$-logP(x|c(x)) = -\sum_i x_i log f_i(c(x)) + (1 - x_i) log(1 - f_i(c(x))) \tag{5.2}$$

The process for training the system is defined by Bengio et al. (2009) as:

1. Train the first layer of the deep architecture as an autoencoder to minimise the reconstruction error from the input

2. The output of the first hidden layer is then used as the input for the next hidden layer, and it is trained as in step (1)

3. Step (2) is iterated until all layers have been trained

4. The output of the last hidden layer becomes the input to a supervised layer (such as logistic regression) and initialises its parameters. Alternatively a final unsupervised reconstruction layer can be placed if the task does not require classification

5. All the parameters of the system are fine-tuned with respect to the output layer (supervised or unsupervised)

As the sAE reconstructs a representation of the original dataset using its decoding stage, certain details of the equalisation parameters will be overlooked. Despite the loss of information being a disadvantage of the process, in the case of equalisation curves it provides a useful tool. Since audio processing is an inherently input dependent task, the sAE acts as a smoothing filter for curves that contain specificities. For example, if a drastic cut was performed to remove noise before changing the timbral characteristic of an audio sample, the sAE will disregard this information as not salient. This means that the system can focus only on the important information that the sAE has found and reproduce these on its decoding stage.

5.3.2 Cluster analysis

The dataset gathered through the SAFE project contains labels regarding the timbral transformations the users are aiming to achieve through audio effects. However, in detecting sub-representations, the task is unsupervised, as users did not define the specific type of *warm* or *bright* they were approximating. Therefore, in order to perform an analysis of their choices, a clustering algorithm will be implemented.

Cluster analysis is an unsupervised learning technique, that is used to group instances of a dataset into categories. Since cluster analysis is agnostic to the data labels, it separates the instances by taking into account the structural information of the dataset, mainly through geometrical measurements, and is widely used for data mining applications (Berkhin et al., 2006).

Clustering techniques

Because of its widespread use in different fields, ease of operation and unsupervised learning attribute, many clustering techniques have been proposed through the years. An early suggestion was the separation of techniques into two main groups, hierarchical and partitioning (Fraley and Raftery, 1998). Hierarchical methods provide a proximity matrix between all points, while partitioning methods, such as *K-means*, separate the space into clusters based on the geometry of the data. In turn, further research has resulted in more methods for cluster analysis, such as density-based, for example *DBSCAN* (density-based spatial clustering of applications with noise) (Ester et al., 1996), distribution-based models,

which include *Gaussian mixture models* (using the expectation-maximization algorithms) and grid-based models, in the *Wavecluster* approach proposed by Sheikholeslami et al. (1998). Special cases of clustering algorithms, outside of the existing categories have also been developed in techniques like *Affinity Propagation*, which uses the 'passing of messages' between data points (Frey and Dueck, 2007), and *Spectral Clustering* (Shi and Malik, 2000), where eigenvalues of a similarity matrix are used to perform the partitioning of data into clusters. In response, Estivill-Castro (2002) criticises this variety of existing techniques, considering it a side-effect of lacking a proper definition of what a "cluster" means, a problem not unlike the various definitions of timbre and its semantic terms. In their review of cluster analysis, Ilango et al. (2011) detail a number of unresolved issues in the discipline. The most significant problems include the optimisation of the clustering algorithm and the evaluation methods used to rate their performance.

Cluster optimisation Clustering techniques can be separated into two further categories: algorithms that require an input number of clusters, and algorithms that require optimisation of parameters other than the number of clusters. *DBSCAN* for example requires the optimisation of two parameters, the maximum distance between two points belonging to a neighbourhood (denoted as *eps*), and *k* the number of points belonging to a neighbourhood for a given point to be considered a "core" point. Ester et al. (1996) suggests that *k* needs to be defined by the researcher without the use of some metric or formula, and in turn the optimal value for *eps* can be detected. Tan et al. (2006) suggest calculating the distances between point p to its *k*-th closest neighbour, and sorting these in ascending order. The point where there is a sharp change in the resulting distance plot will pinpoint the value for the *eps* parameter. Similarly, *Spectral Clustering* requires a different heuristic approach to identify the number of input clusters. By computing the eigenvalues of a Laplacian matrix, von Luxburg (2007) suggests detecting the *eigengap*, a point of sudden change between the eigenvalues, that will indicate the appropriate number of clusters. There exists one main issue regarding the methods in which cluster parameters are derived, which is that the measurements depend on essentially arbitrary techniques instead of established metrics. The identification of the *eigengap* for *Spectral Clustering* is not clearly defined, as it is not clear when an *eigengap* is present. Similarly, for *DBSCAN* there does not exist a clear definition for the sharp change found in the distance plot. This results in the decision for the

parametrisation of the techniques being made by the researcher's definition of what should be considered an *eigengap* or a sharp change, and therefore the choice is subjective.

Cluster evaluation Evaluating clusters can be performed in a supervised or in an unsupervised manner. Since the application features an unsupervised task, unsupervised evaluation methods are implemented. Here, the standard approach according to Halkidi et al. (2002) is to measure (a) how rigid and condensed the data in a cluster are (*cohesion*), and (b) how remote one cluster is from the others (*separation*). As a result, most unsupervised evaluation metrics focused on combining cohesion and separation. These include techniques such as the Davies-Bouldin index (Davies and Bouldin, 1979), the Dunn index (Dunn, 1973), the silhouette metric (Rousseeuw, 1987) and comparing the correlation of the similarity matrix with an 'ideal' version of itself (Tan et al., 2006). However, an inherent problem of these metrics is that the error produced will continue to minimise as the number of clusters is increased. This is because a higher number of clusters will increase the *cohesion* of each group, while maximising the overall *separation* between clusters. Additionally, some metrics used for evaluation, as the silhouette coefficient, are also used for finding the optimal number of clusters, for techniques like *K-means* (Kodinariya and Makwana, 2013).

Technique comparison In order to compare clustering techniques, the number of clusters must be set to the same value. This is a trivial problem for techniques that allow for an input number of clusters, such as *K-Means* or *Spectral Clustering*, but it becomes more complicated when these need to be compared to techniques such as *DBSCAN*, which requires parameter optimisation. In addition, there is no guarantee that the optimal number of clusters will be the same for all techniques applied to a given dataset. Similarly, the unsupervised metrics will display a favourable result to techniques that include a higher number of clusters.

Due to these details, an alternative method for choosing, optimising and evaluating clustering is presented in Section 5.4.2.

5.4 Data Preparation

5.4.1 Intrinsic Dimension

Whenever a dimensionality reduction technique is applied, the algorithm will disregard certain information, which it finds to be not salient. However, it is essential that salient information is retained in the reconstruction process. Therefore, it is necessary to identify a method for choosing the parameters of the sAE model.

Considering the most widely used dimensionality reduction technique in PCA, the number of dimensions for the low-dimensional mapping chosen by researchers will be dependent on the variance explained in the system. As Ferrando and Lorenzo-Seva (2013) explain, a number of principal components needs to be chosen so that it can adequately describe the variance inherent in the original high-dimensional dataset.

However, as not all techniques can be optimised by such methods, Van Der Maaten et al. (2009) considers the existence of a number of dimensions below which the loss of information surpasses an acceptable threshold. This concept is known as the intrinsic dimensionality of the dataset, defined by Kalantan (2014) as the minimum number of components/dimensions that are necessary to describe the data without significant loss of information. This in turn would mean that reducing the number of dimensions below the intrinsic dimension would result in the low-dimensional representation being inaccurate, or the reconstruction being poor. Levina and Bickel (2005) have proposed a technique for uncovering the intrinsic dimension based on a maximum likelihood estimation method, and this will be used to measure the intrinsic dimensions of the *warm* and *bright* datasets.

In computing the intrinsic dimension of the two datasets (*bright* and *warm*), by using the maximum likelihood estimation method it is found that both datasets can be reduced without significant information loss to two dimensions. Therefore, for the implementation, a sAE of two hidden layers is used, having sizes [13 - 9 - 2], with the network optimised using stochastic gradient descent. As the sAE features a mirrored architecture, the 13 parameters will be reduced to a number of 2, which will then be used to provide a reconstruction of the original 13. In that process any information that is not considered important by the system will be disregarded. This was implemented using the Theano Python library (Bergstra et al., 2010).

5.4.2 Cluster analysis

Following the application of the sAE model, in order to partition the resulting space, cluster analysis needs to be performed. This process will allow for the identification of the sub-representations of single descriptive terms. To partition the data, the methodology presented by Tan et al. (2006) is followed. Here, the five key steps are:

- Identification of the clustering tendency in the dataset
- Detection of the optimal number of clusters / parameter optimisation
- Evaluation of the algorithm without reference to external information
- Evaluation of the algorithm with reference to external information
- Comparison of multiple clustering algorithms to identify the most suitable method

This splits the evaluation section into three unsupervised tasks, a supervised evaluation task and a comparative process. Since the data does not contain labels regarding the sub-representations, the supervised task is excluded, and due to the comparison problems between techniques addressed in Section 5.3.2, a comparative analysis between different algorithms is not implemented. Instead for the final evaluation of the system and the identification of sub-representations, listening tests will be performed.

The chosen methodology is presented in Figure 5.5, and will include three objective evaluation steps in identifying the clustering tendency, choosing and optimising the clustering algorithm and finally evaluating its performance. Due to the type of clustering technique used in this research, the clustering evaluation can precede the clustering optimisation step. Finally, in order to take advantage of relevant metrics, the EQ parameters are converted to EQ curves (as was depicted in Figure 5.3).

Clustering tendency

Clustering can be performed on a dataset without taking into account its structural characteristics. However, before any clustering is applied, it is necessary to measure the extent to which clusters may exist in the dataset.

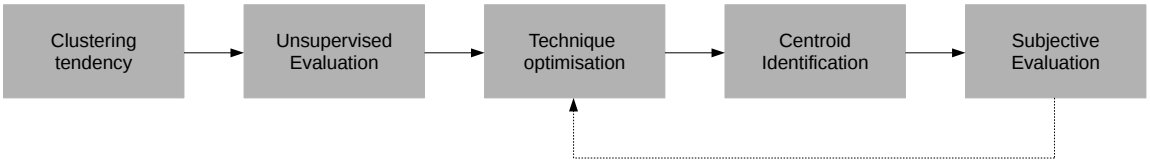


Figure 5.5: Cluster analysis methodology for uncovering the salient sub-representations of *warm* and *bright*.

Clustering tendency is defined as the extent to which the structure of a dataset includes well-defined clusters, also known as natural clusters. A popular approach in measuring the clustering tendency is through statistical testing of the spatial randomness (Tan et al., 2006), or by using the Hopkins Statistic (Hopkins and Skellam, 1954).

In order to calculate the Hopkins Statistic, p points (Set 1) randomly distributed in the data space are generated, and p points (Set 2) are sampled from the existing dataset. The following step is to measure the distance of the points in both sets to their nearest neighbour, where dr is the distance of the sampled points to their nearest neighbour and di is the distance of the generated points to their nearest neighbour over n dimensions. The Hopkins Statistic is then calculated as:

$$H = \frac{\sum_{i=1}^{p} di^n}{\sum_{i=1}^{p} dr^n + \sum_{i=1}^{p} di^n} \tag{5.3}$$

The formula will produce a number in a range from 0.5 to 1, where a high value represents a highly clustered dataset, and a low value suggests the data is randomly distributed in the space.

For the *bright* descriptor, over a 20-fold cross-validation the Hopkins Statistic produces a mean value of 0.54388 with a standard deviation of 0.027, and for the *warm* descriptor a mean value of 0.561 with a standard deviation of 0.043. From these results it can be deduced that the clustering tendency of the two descriptors is similar, suggesting a common arrangement of the data in the space. Furthermore, the Hopkins statistic measurement suggests that the datasets are randomly distributed in the space, however that does not necessarily mean that sub-sets of the two descriptors do not exist in these spaces. The metric is designed to detect natural clusters, groups of points that exhibit high intra-class cohesion and high between-class separation. This suggests that clusters can still exist in the *warm* and *bright* spaces, but their presence is not apparent.

Choice of clustering technique

In her criticism of clustering techniques, Estivill-Castro (2002) suggests that the choice of the clustering algorithm for a given task should be made based on the research topic in question. This is proposed because of the data dependent performance of clustering algorithms, where a technique might be optimal for task A, but will not display the same accuracy for task B. Additionally, Ilango et al. (2011) indicate that another common problem in cluster analysis is the definition of an appropriate distance metric. These two obstacles can be bypassed by implementing an algorithm that is informed of the research at hand, and will make use of a metric relevant to the entries in the dataset.

A variation of hierarchical clustering is therefore proposed as the optimal approach for uncovering salient subsets of equalisation curves. Taking advantage of the ability to define a metric for hierarchical clustering, the modified algorithm aims at partitioning the data based on their spectral representations. By implementing this modified distance function to the algorithm, the system takes advantage of knowledge regarding the specific application. The metric applied is based on the coherence measurement, defined in Equation 4.29 from Section 4.3.3.

By looping through all the combinations of instances, a distance matrix is constructed by finding the pairwise coherence measurements between all frequency bins of the reconstructed equalisation curves. Through the implementation of this modified metric it is expected that the resulting clusters will share characteristics based on a criterion that identifies the level of similarity between their frequency distributions, rather than their structure and geometry in the given space.

Additionally, as was performed in Section 4.3.3, the results of the coherence metric are reversed and rescaled in order to represent them as distance measurements.

Clustering performance evaluation

As hierarchical clustering does not need an input number of clusters to be defined the evaluation process can be performed first, and then the optimal linkage will be used to find the appropriate number of clusters.

Specifically in the case of hierarchical clustering, it is possible to use an evaluation metric on the performance of the linkage option. As hierarchical clustering constructs a

hierarchy of the input dataset in the form of dendrograms, a common practice in evaluating the performance of the algorithm is through the cophenetic distance (Lapointe and Legendre, 1995). Since dendrograms are graphical representations of a cophenetic matrix, it is possible to identify the optimal linkage technique for the algorithm through the cophenetic correlation. This is defined as:

$$c = \frac{\sum_{i<j}(x(i,j) - x)(d(i,j) - d)}{\sqrt{[\sum_{i<j}(x(i,j) - x)^2] + [\sum_{i<j}(d(i,j) - d)^2]}} \tag{5.4}$$

where $x(i,j)$ is the Euclidean distance between points i and j, and $d(i,j)$ is the distance between the two points in the dendrogram. x and d represent the average values of $x(i,j)$ and $d(i,j)$.

In order to detect the optimal linkage method, the cophenetic correlation of all linkage options is calculated (Equation 5.4). The seven types of linkage that are evaluated include: single, complete, average, weighted, centroid, median and Ward, and Table 5.1 displays their performance.

Linkage	***Bright* Cophenetic Correlation**	***Warm* Cophenetic Correlation**
Single	0.320214	0.522341
Complete	0.862697	0.685418
Average	**0.878668**	**0.812211**
Weighted	0.800794	0.703954
Centroid	0.870976	0.811222
Median	0.846046	0.682768
Ward	0.862679	0.675793

Table 5.1: Comparison of the cophenetic correlation across different linkage options for both descriptors. Higher values suggest the linkage choice is able of retaining the distance relationships between data entries.

For the *bright* descriptor, the average method performs with higher accuracy, while centroid, complete and Ward, produce slightly less accurate results. The *warm* descriptor exhibits a similar behaviour, with the average type scoring the highest, while centroid linkage performs slightly less favourably. The results of this evaluation are in accordance with a comparative study on the performance of linkage metrics (Saraçli et al., 2013), where for a set of artificially generated datasets, average and centroid linkages performed best.

Optimisation of clustering parameters

Hierarchical clustering does not make use of arbitrary methods for identifying the optimal number of clusters, as is the case with *DBSCAN* and *Spectral Clustering*, and it is not optimised through a metric that can also be used for evaluating its performance, as with *K-Means*. For hierarchical clustering, the inconsistency metric can be implemented to detect the appropriate number of clusters. This is defined by Martinez et al. (2010) as:

$$inconsistency = \frac{h - avg}{std} \tag{5.5}$$

where h is the height of the links, avg is the mean of the heights of all links, and std is the standard deviation of all links.

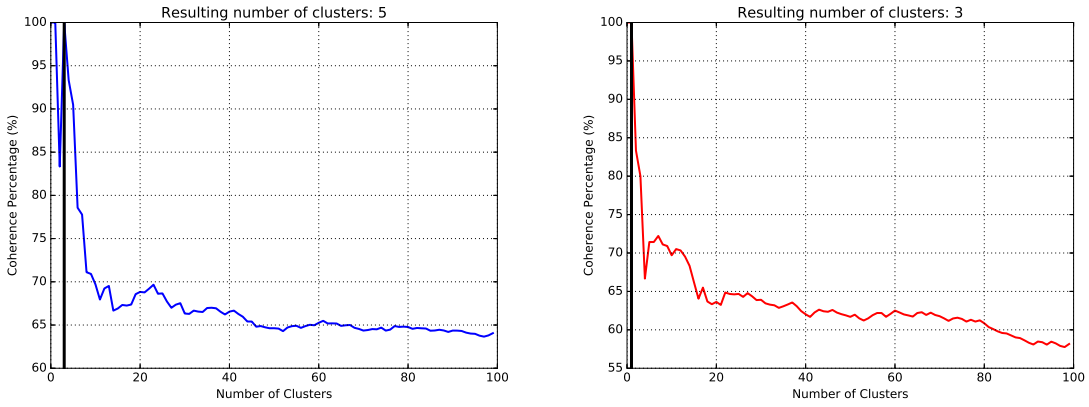
However, a depth parameter needs to be defined in order to determine the number of links to be examined. Due to its dependency on the depth parameter, the inconsistency method for finding the optimal number of clusters produces varying results. As is apparent from Table 5.2 for both *warm* and *bright*, increasing the depth value will result in a higher number of clusters. In the case of the *bright* descriptor the result varies from two to eleven clusters, while for *warm* from two to ten.

Dataset	*Depth*	*Inconsistency Coeff*	*Number of Clusters*
Bright	1	0.0	1
	2	1.06	2
	3	2.04	3
	4	3.09	6
	5	4.42	11
Dataset	*Depth*	*Inconsistency Coeff*	*Number of Clusters*
Warm	1	0.0	1
	2	1.03	2
	3	1.908	3
	4	2.909	6
	5	4.05	10

Table 5.2: Inconsistency measurement with varying values of depth for detecting the optimal number of clusters. From the table it can be shown that the depth parameter affects the number of resulting clusters.

Due to the dependency of the inconsistency formula on the depth parameter, an alternative for detecting the optimal number of clusters is considered. Hierarchical clustering is performed iterating the number of clusters (N_c) from 2 to 100, and on each iteration the

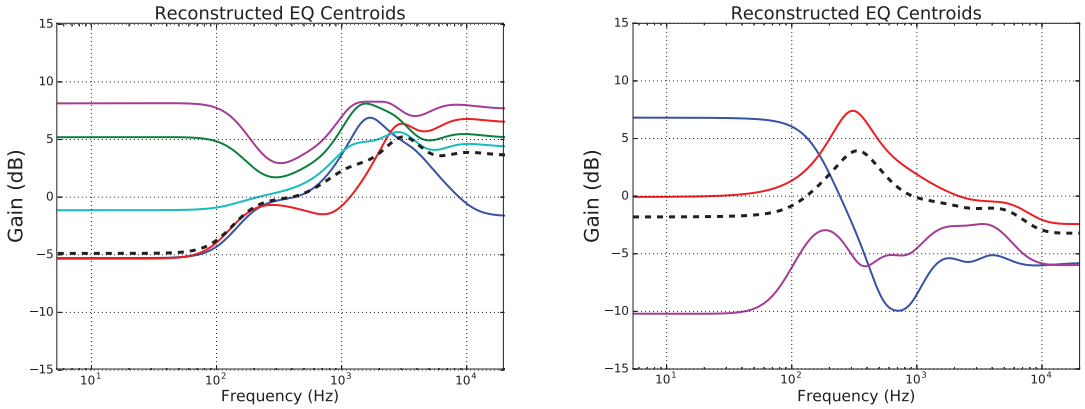
coherence is measured between the cluster centroids. By empirically setting a threshold for the coherence of +0.7, a value that generally suggests a strong positive correlation, a value of 1 is assigned to the pair if it has a coherence greater than the threshold, and a value of 0 otherwise. A percentage measurement is then calculated based on the number of pairs that did not meet the threshold criteria. The point where N_c is found to be largest and possess no clusters that surpass the threshold is considered optimal. This is displayed as the vertical black lines in Figures 5.6a and 5.6b.



(a) (b)

Figure 5.6: Coherence ratings for (a) *bright* and (b) *warm* descriptors over varying number of clusters. The black vertical line signifies the highest number of clusters where the coherence between all sub-representations is below the set threshold of +0.7.

The process of finding the optimal number of clusters, where all measurements between centroids fall below the +0.7 threshold is depicted in Figure 5.6. The resulting number is $N_c = 3$ for the *warm* descriptor, and $N_c = 5$ for the *bright* descriptor. The coherence matrices featuring the relationship between all cluster centroids are presented in Tables 5.3 and 5.4 for *bright* and *warm* respectively, and the resulting equalisation curves for each sub-representation are presented in Figure 5.7b for *warm* and Figure 5.7a for *bright*.



(a) (b)

Figure 5.7: The resulting EQ curves sub-representations for (a) *bright* and (b) *warm*, with the dashed line displaying the overall centroid of the term.

Sub-representations	*Bright 1*	*Bright 2*	*Bright 3*	*Bright 4*	*Bright 5*
Bright 1	1.0	-0.328	-0.044	0.58	0.283
Bright 2	-0.328	1.0	0.587	0.073	0.695
Bright 3	-0.044	0.587	1.0	0.666	0.673
Bright 4	0.58	0.073	0.666	1.0	0.67
Bright 5	0.283	0.695	0.673	0.67	1.0

Table 5.3: Coherence distance matrix for *bright* sub-representations.

Sub-representations	*Warm 1*	*Warm 2*	*Warm 3*
Warm 1	1.0	0.108	0.027
Warm 2	0.108	1.0	0.541
Warm 3	0.027	0.541	1.0

Table 5.4: Coherence distance matrix for *warm* sub-representations.

5.5 Subjective evaluation

Figure 5.8 presents the interface for the subjective evaluation process. In this case, the subjects were presented with a number of sliders, each corresponding to an audio sample equalised using a different sub-representation of the semantic term (i.e. *warm* or *bright*). One of the sub-representation samples is chosen as the hidden reference for each comparison, and the participants had to rate the similarities of each sub-representation to the hidden reference.

For example, when the *Bright 1* sample is set as the reference all of the *bright* sub-representations will be compared to it, including the reference. This process was performed for the five sub-representations of *bright*, where subjects had to rate the similarity of five samples compared to a hidden reference, and the three sub-representations of *warm*, with subjects having to rate the similarity of three samples compared to a hidden reference. In the context of this process an anchor sample was not used. This was performed as an anchor sample would have to take on the reverse characteristics of the descriptive term, the choice for which would have to be subjective. In addition, the subjective evaluation needed to be between sub-representations of the same descriptor and therefore it was considered that another descriptive term should not be included alongside these.

The five sub-representations of *bright* and the three sub-representations of *warm* are applied to each of the two audio samples, giving a total of sixteen samples being presented to the listening test subjects. By then summing the subjects' responses, it is possible to investigate the perceptual similarity between audio samples. Figures 5.9a, 5.9b, 5.10a and 5.10b display the subjects responses, corresponding to each reference (top of figures). For instance, in Figure 5.9b the elements of the legend allocated on the top of the figure (EQ1, EQ2, EQ3) correspond to the reference of each MUSHRA test. The first three bars display the subjects' similarity scores when compared to EQ1, where it can be seen that the sample belonging to the same sub-representation as the reference (EQ1) achieves the highest similarity score, while the two remaining sub-representations display a lower degree of similarity to the reference. The standard deviation is also included in the plot, displayed by the thin colored line appended to each bar. This measurement can aid in detecting the subjects' confidence in each response. For instance, in Figure 5.9b for the third MUSHRA test (EQ3), the correct response also features a shorter error bar than the correct response for

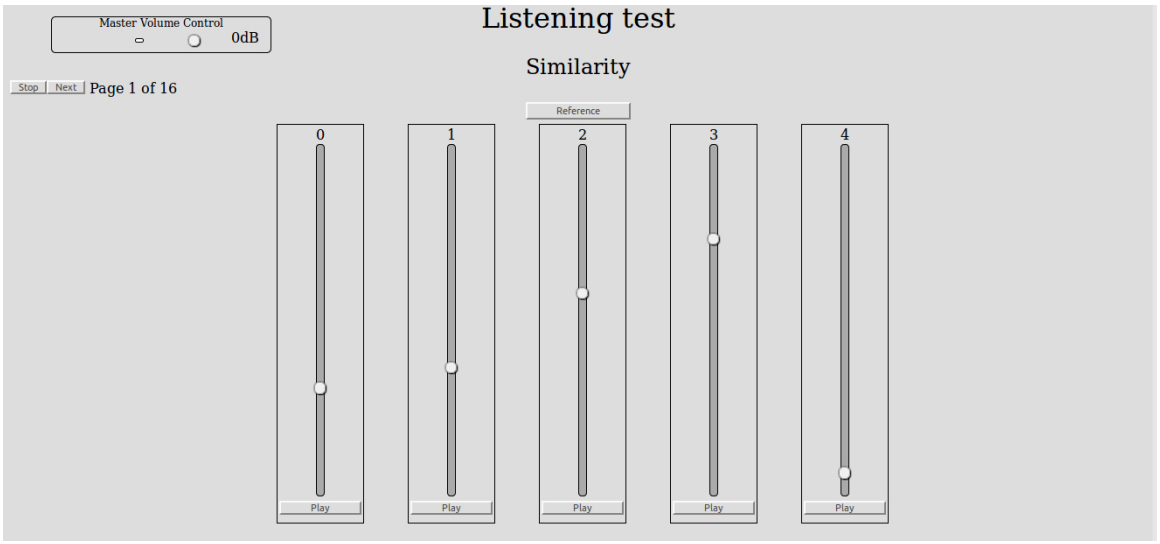
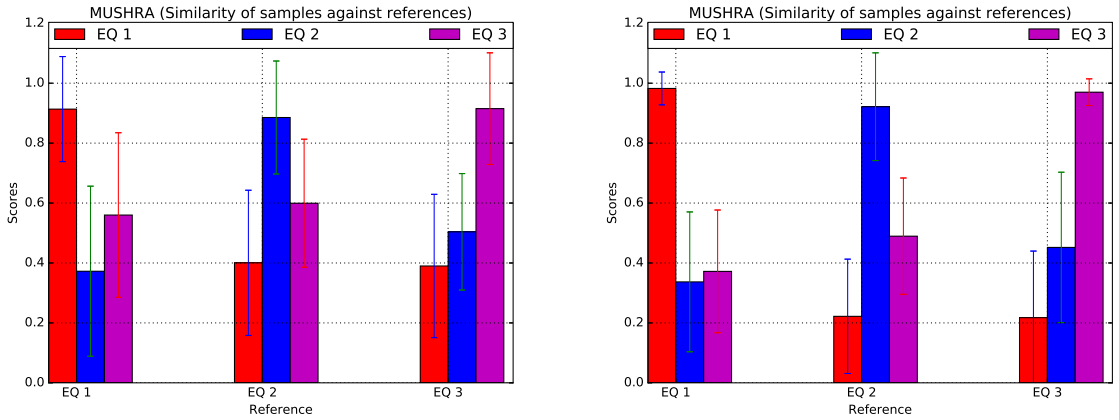


Figure 5.8: Snapshot of the MUSHRA listening test interface. Here the vertical sliders were used to rate the similarity of the individual sub-representations to the reference sample (top middle toggle).

the second MUSHRA test (EQ2), suggesting that the subjects were more confident about their choice in picking the appropriate response for EQ3 than when making a choice for EQ2.

The first question posed by this research is whether the test subjects will be able to detect the correct sub-representation in a MUSHRA environment. As displayed in Figure 5.8, the subjects were asked to rate the similarity between samples to a reference. In this task the subjects were successful, allocating the reference sample to the correct sub-representation for the *warm* descriptor across both genres, as is displayed in Figures 5.9a and 5.9b. For the *bright* descriptor, subjects were also able to detect the reference to the correct sub-representation for the *metal* sample, depicted in Figure 5.10b, but there were some deviations for the *blues* sample (Figure 5.10a). The only incorrect allocation is found in the case of the *blues* samples processed with the *bright* sub-representations, where subjects correctly allocated four out of the five equalisation curves. This brings the total prediction for the *bright* descriptor down to 87.5%, and the total across descriptors to 93.75%. This result is expected, as certain sub-representations for the *bright* descriptor display strong positive correlation values, between +0.66 and +0.695 (Table 5.3), while the *warm* descriptor exhibits less strong correlation between its sub-representations (Table 5.4).

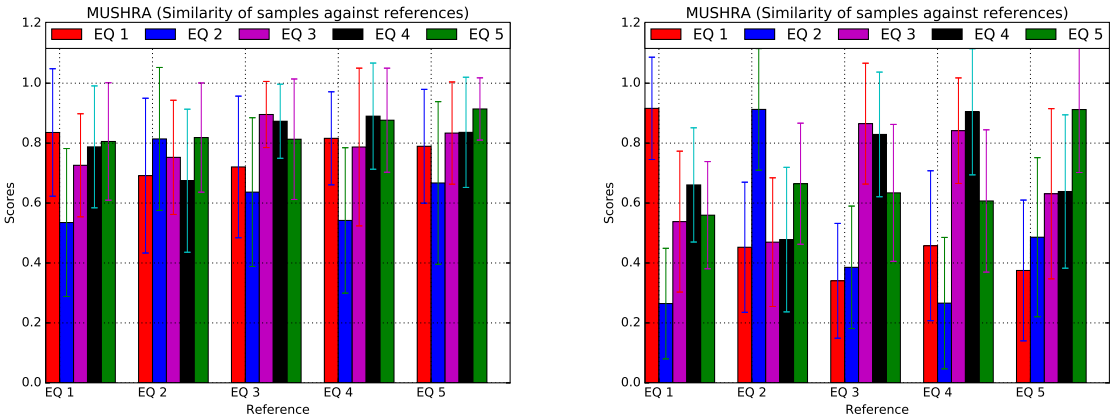


(a) (b)

Figure 5.9: Barplot of the MUSHRA results for the (a) Blues *Warm* and (b) Metal *Warm* samples for the three sub-representations of the term. The legend at the top of the figures depicts the reference for each MUSHRA test.

Additionally, by calculating the standard deviation of the subjects' responses, it is possible to measure the confidence in their choices. As is shown in Figures 5.9a and 5.9b, the correct response also displays the lowest standard deviation. This suggests a higher level of confidence from the side of the subjects in choosing the correct sample for the *warm* sub-representations. This behaviour does not translate to the subjects' responses regarding the *bright* sub-representations, where the listeners present a lower level of confidence for their choices. In more detail, for both the *blues* (Figure 5.10a) and the *metal* (Figure 5.10b) guitar samples, the subjects exhibit the lowest confidence on the correct answers, for two out of the five tests.

Finally, in order to measure the extent to which the pairwise coherence distances are preserved in the perceptual testing, the Spearman correlation (Equation 4.27) is calculated. Both the coherence matrices (Tables 5.3 and 5.4) and the subjective responses (Figures 5.10 and 5.9) are converted into ranks, and it is found that for the *warm* descriptor the Spearman score is +0.83, while for the *bright* descriptor a score of +0.79 is achieved. This suggests that the coherence-based hierarchical clustering algorithm is capable of partitioning the dataset appropriately, and retain the perceptual structure between sub-representations.



(a) (b)

Figure 5.10: Barplot of the MUSHRA results for the (a) Blues *Bright* (b) Metal *Bright* samples for the five sub-representations of the term. The legend at the top of the figures depicts the reference for each MUSHRA test.

5.6 Discussion

5.6.1 Fine-tuning the algorithm for the *bright* sub-representations

As high positive correlation suggests there exist significant similarities between the *bright* EQ curves, this could explain the difficulty exhibited by subjects in recognising the correct audio samples in the subjective evaluation task. It is therefore considered that a lower number of clusters might surpass some of these problems. Since all the pairwise coherence measurements for the *bright* descriptor fall twice below the threshold through the iterations, a different number of clusters can also be set. Figure 5.6a depicts the measurement reaching an acceptable result for $N_c = 3$ as well as $N_c = 5$. The resulting relationship between all cluster centroids is presented in Table 5.5 and the final *bright* and *warm* sub-representations are shown in Figure 5.11.

The need for an additional subjective evaluation processing step after fine-tuning the system is bypassed. Hierarchical clustering does not perform partitioning of the dataset in clusters containing an equal number of instances (as K-means does). In this case three of the original five *bright* sub-representations were merged to form one sub-representation. The relationship between the first and second *bright* sub-representations, achieving a score

of -0.328, is unaltered from this process. Therefore these do not need to be tested again. Furthermore, the relationships between the new sub-representation (*Bright 3*) to the two previous sub-representations displays positive correlation ranging from 0.317 to 0.535. This can be considered an acceptable range for subjects to derive perceptual differences between the EQ curves, as they have performed for a similar range for the *warm* descriptor, where coherence scores ranged from 0.27 to 0.541.

Sub-representations	*Bright 1*	*Bright 2*	*Bright 3*
Bright 1	1.0	-0.328	0.317
Bright 2	-0.328	1.0	0.535
Bright 3	0.317	0.535	1.0

Table 5.5: Coherence distance matrix for the three final *bright* sub-representations.

Having fine-tuned the resulting sub-representations, these will now be named based on their characteristics on the frequency spectrum. The first *warm* sub-representation presents a boost on the low-end that takes the shape of a shelving filter and therefore will be called low-shelf boost (LSB) *warm*. The second *warm* sub-representation is depicted with a boost on the low-mid range, and will be called low-mid boost (LMB) *warm*. Finally, the third warm sub-representation displays a higher cut-off in the low range, while most energy is concentrated on the mid and high-mid frequencies. This will be called high-mid boost (HMB) warm.

Similarly, the *bright* sub-representations will be named after the characteristics of the EQ curves. The first *bright* sub-representation displays a high-mid range boost, and will be called high-mid boost (HMB) *bright*. The second sub-representation features a high-end shelving boost and a low-shelf cut, and will be called high-shelf boost (HSB) *bright*. Finally, the third bright sub-representation displays a slight low-shelf boost and also a high-shelf boost, and it will be called low-shelf boost (LSB) *bright*.

5.6.2 Cross-term sub-representation correlation

Propositions on the relationship between *warm* and *bright* descriptors vary in existing research, with Brookes and Williams (2007, 2010) suggesting that each term can be altered independently of the other, while Zacharakis and Reiss (2011) propose that there is an an-

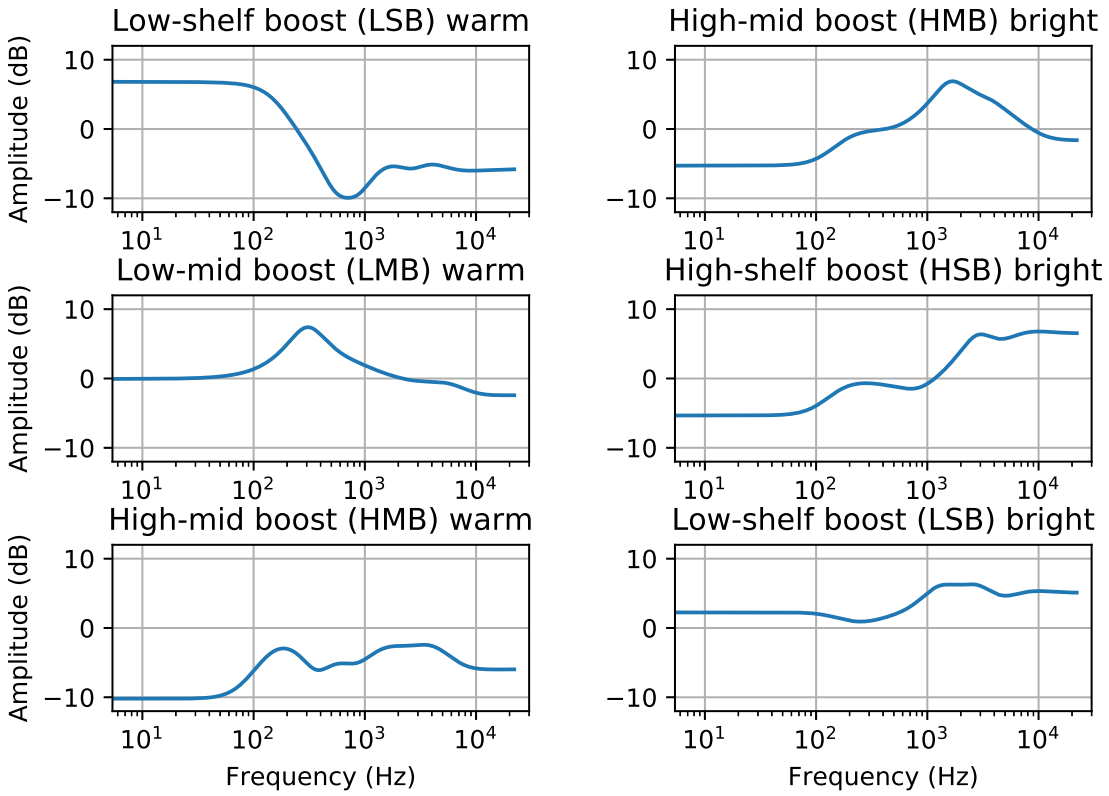


Figure 5.11: Final sub-representations for the *bright* and *warm* descriptors.

tithesis between the two. Based on this, the coherence (Equation 4.29) between the salient sub-representations of the two terms can be measured, to provide an extent of the relationship between the two descriptors.

As Table 5.6 displays, the relationship between *warm* and *bright* is altered depending on the sub-representation measured. LSB *warm* displays a coherence ranging from -0.1 to -0.33, suggesting little to no correlation with the *bright* sub-representations. LMB *warm* sub-representation displays a more interesting behaviour, achieving a strong negative correlation with HSB *bright* (-0.84) and a weaker negative correlation (-0.53) with LSB *bright* sub-representation. However, it also displays a weak to strong correlation with HMB *bright* (+0.57), which can be explained by the similar amplification of the mid frequencies present in both curves. On the other hand HMB *warm* sub-representation, while displaying weak

Sub-representations	*LSB warm*	*LMB warm*	*HMB warm*
HMB bright	-0.103	0.571	0.943
HSB bright	-0.197	-0.841	-0.231
LSB bright	-0.326	-0.53	0.218

Table 5.6: Coherence measurements between *warm* and *bright* sub-representations.

correlation with HSB (-0.23) and LSB (+0.22) *bright*, achieves a very high score for the HMB *bright* (+0.94).

The majority of the coherence measurements fall either in the negative correlation range, or exhibit a correlation close to 0, suggesting that for achieving the two descriptors through equalisation, opposing or independent settings need to be applied. However, there is also a strong connection regarding HMB *bright* with the LMB *warm* and, especially, HMB *warm*, which suggests there is also a point where equalisation parameters converge with regard to *warm* and *bright* settings.

5.6.3 Correlation with additional descriptors

As the SAFE EQ dataset presents a wide range of different terms, their similarity to each of the *warm* and *bright* sub-representations can be compared. It is therefore possible to reveal the multiple definitions that *warm* and *bright* possess, where individuals will use different terms to describe them. Table 5.7 shows the correlation scores between the three *bright* sub-representations to the other terms in the dataset. It can be seen that the first sub-representation, HMB *bright*, correlates highly with the terms *tin, full, bass, mid, thick* and *bite*. The second sub-representation, HSB *bright*, displays a high coherence rating with the terms *crisp, presence, air* and *cut*. Finally, the third bright sub-representation, *LSB bright*, achieves a high coherence rating with the term *bite*.

Similarly, Table 5.8 shows the correlation scores of the three *warm* sub-representations to the other terms of the SAFE EQ dataset. It is found that the first sub-representation, LSB *warm*, does not correlate with any other terms, apart from a moderate positive relationship with *boost*. The second sub-representation, LMB *warm* is found to strongly correlate with the terms *deep, boom, mid, bass, thick, box* and *full*. Finally, the third sub-representation, HMB *warm* correlates strongly with the terms *tin, bass, full, mid, thick* and *bite*.

Sub-representations	*HMB bright*	*HSB bright*	*LSB bright*
bite	0.724	0.314	0.753
air	-0.565	0.82	0.44
tin	0.96	-0.126	0.484
click	0.062	0.455	0.659
cut	-0.517	0.776	0.262
clean	0.39	0.164	0.374
thin	0.383	0.442	0.632
mid	0.874	-0.582	-0.122
presence	-0.123	0.908	0.369
clear	0.39	0.164	0.374
thick	0.853	-0.433	-0.095
crisp	-0.184	0.909	0.691
mud	0.312	-0.724	-0.03
low	0.313	0.475	0.426
vocal	0.195	-0.363	0.113
bass	0.927	-0.583	0.004
box	0.426	-0.85	-0.198
boom	0.578	-0.89	-0.445
punch	0.446	0.264	0.067
boost	-0.201	-0.68	-0.803
full	0.935	-0.445	0.076
deep	0.464	-0.92	-0.604

Table 5.7: Coherence measurements between *bright* sub-representations to other popular terms from the SAFE EQ dataset.

Figure 5.12 presents a hierarchy of the terms regarding their relationship to the *warm* and *bright* sub-representations, which re-affirms their categorisation. The *bright* sub-representations are split into two categories, with HSB *bright* and LSB *bright* displaying an overlap on both the terms they are positively and negatively correlating with, while the HMB *bright* sub-representation displays a similarity with descriptors usually not associated with brightness in *tin, bass, full, boom* and *thick*. Similarly, the *warm* sub-representations are also split into two groups. LMB *warm* and HMB *warm* sub-representations display correlations with terms such as *mid, bass, full* and *thick*. Interestingly, the first sub-representation (LSB *warm*), as it does not display any strong correlation with other terms, can be considered a standalone definition of warmth.

Sub-representations	*LSB warm*	*LMB warm*	*HMB warm*
bite	-0.127	-0.054	0.716
air	-0.235	-0.916	-0.486
tin	-0.154	0.379	0.918
click	-0.334	-0.383	0.041
cut	-0.24	-0.744	-0.373
clean	-0.203	-0.068	0.286
thin	-0.812	-0.225	0.327
mid	0.139	0.8	0.862
presence	-0.099	-0.582	0.034
clear	-0.203	-0.068	0.286
thick	-0.197	0.745	0.84
crisp	-0.444	-0.836	-0.143
mud	-0.397	0.495	0.122
low	-0.247	-0.197	0.313
vocal	-0.469	0.32	0.054
bass	0.07	0.784	0.907
box	-0.291	0.702	0.253
boom	0.221	0.952	0.553
punch	0.197	0.05	0.599
boost	0.511	0.525	-0.212
full	-0.081	0.702	0.867
deep	0.206	0.972	0.446

Table 5.8: Coherence measurements between *warm* sub-representations to other popular terms from the SAFE EQ dataset.

5.6.4 Correlation with existing literature

Chapter 4 presented frequency responses generated using a number of timbral adjectives gathered from the existing literature (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005) by De Man (2017). Through the individual definitions for the *bright* and *warm* terms, it is possible to examine the extent to which the found sub-representations correlate with the different interpretations used by professional sound engineers.

As illustrated in Table 5.9, for the *warm* sub-representation, LSB *warm* displays a moderate to strong correlation with the definitions presented by Owsinski (2013) (+0.6333) and Huber and Runstein (2013) (+0.474). On the other hand, it displays a weak negative correlation with the definition of Gibson (2005) (-0.256). LMB *warm* displays varying degrees of positive correlation with all the heuristic definitions (Owsinski, 2013; Izhaki, 2013; Hu-

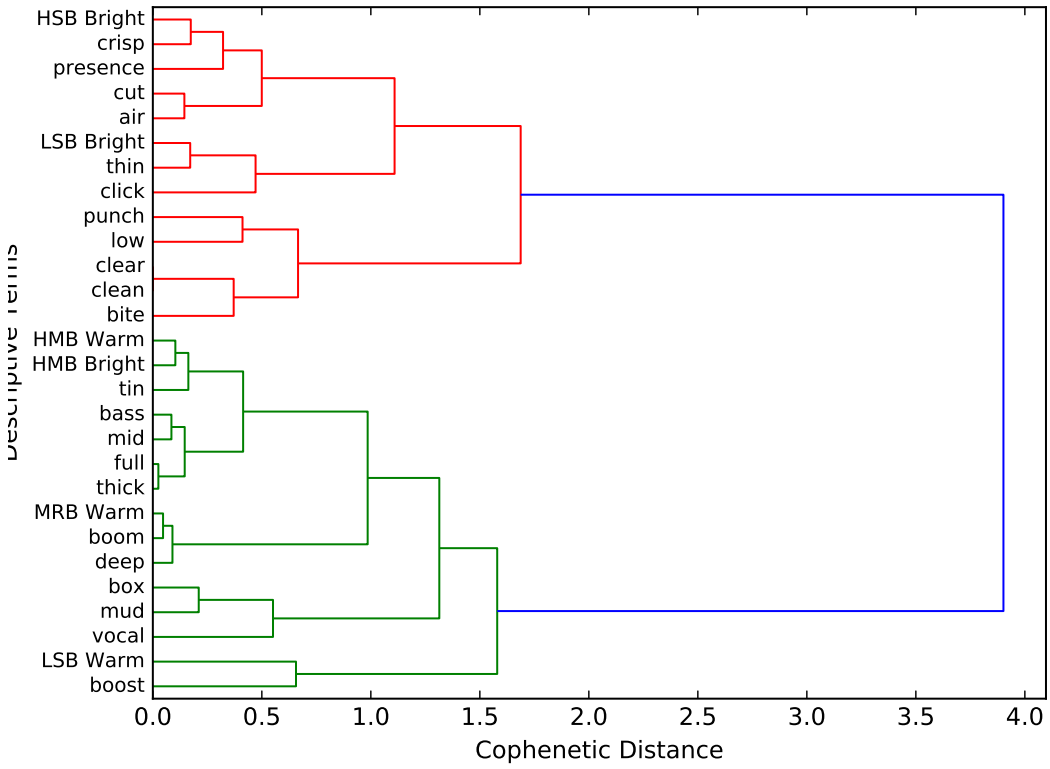


Figure 5.12: Hierarchy of terms with regard to their relationship with the *warm* and *bright* sub-representations.

ber and Runstein, 2013; Gibson, 2005; Cousins and Hepworth-Sawyer, 2013; Katz, 2003), ranging from +0.349 to +0.862. Finally, HMB *warm* displays very weak to no correlation with the heuristic definitions, ranging from +0.048 for the definition provided by Cousins and Hepworth-Sawyer (2013) to +0.148 for the one attributed to Gibson (2005).

While there appears to be some agreement regarding practical sub-representations of the *warm* descriptor, the same behaviour is not found for the *bright* sub-representations. By comparing the resulting sub-representations with the heuristic definitions gathered by De Man (2017), the coherence measurements are illustrated in Figure 5.10. Here, it is found that HMB *bright* exhibits a weak positive correlation with both the definitions provided by Katz (2003) and Gibson (2005), +0.233 and +0.228 respectively, while also displaying a weak negative correlation with the definition attributed to Izhaki (2013). HSB *bright*

Heuristic Definitions / Sub-representations	*Owsinski*	*Izhaki*	*Huber*	*Gibson*	*Cousins*	*Katz*
LSB warm	0.633	-0.002	0.474	-0.256	0.035	-0.14
LMB warm	0.349	0.862	0.647	0.82	0.787	0.776
HMB Warm	0.055	0.119	0.095	0.148	0.048	0.051

Table 5.9: Coherence measurements between the *warm* sub-representations and heuristic definitions.

also achieves a weak positive correlation with the definitions of Gibson (2005) (+0.134) and Katz (2003) (+0.274), while also reaching a moderate positive correlation with the definition of Izhaki (2013) (+0.449). However, LSB *bright*, appears to have little to no correlation with these heuristic guidelines (ranging from -0.074 to 0.153).

Heuristic Definitions / Sub-representations	*Katz*	*Izhaki*	*Gibson*
HMB bright	0.233	-0.319	0.228
HSB bright	0.274	0.449	0.134
LSB bright	0.059	0.153	-0.074

Table 5.10: Coherence measurements between the *bright* sub-representations and heuristic definitions.

5.7 Conclusion

5.7.1 Overview

This chapter has presented a method for categorising definitions of the same term into different groups to identify divergent definitions of the same semantic term. A modified version of hierarchical clustering has been implemented, setting a strict threshold on the resulting groups. In turn subjects have found that the three *warm* sub-representations possess unique characteristics and are distinguishable from each other, while for the original, five *bright* sub-representations perceptual overlap occurred. For this reason the *bright* sub-representations were reduced to a number of three instead, to preserve their individual characteristics.

Additionally, by measuring the coherence across terms, it is found that different in-

terpretations of each descriptor will affect its relationship with the other. It is shown that whilst the global definitions of *bright* and *warm* were found to be antonymous in Section 4.4.3 (Table 4.8), the sub-representations of each will produce varied results. As presented in Table 5.6 these can range from high negative correlation (-0.841) to high positive correlation (+0.943).

Furthermore, a hierarchy of correlating terms was produced between all sub-representations of *warm* and *bright*, revealing a mostly opposing characteristic between the two, but also a point of overlap. In more detail, the *bright* descriptors appear to encompass the terms *bite, click, thin and crisp*, and to a lesser extent the terms *air* and *tin*. The *warm* descriptor on the other hand corresponds with the terms *tin, mid, thick, boom, full* and to a lesser extent *deep* and *boost*. In addition, the relationship of the sub-representations with heuristic definitions of the terms has been examined, concluding that *warm* for the most part exhibits a positive correlation with the heuristic definitions, while *bright* sub-representations exhibit a slight deviation from these guidelines.

Finally, this chapter introduced the sAE model as a method for attaining salient information from the dataset. Considering the architecture of the sAE, by which the dataset is first reduced to a lower number of dimensions and then approximately reconstructed, it is possible to use this process as a basis for constructing novel interfaces for intelligent audio effects. The following chapter will expand on the use of dimensionality reduction and parameter reconstruction methods for building such systems.

5.7.2 Limitations

While this study has provided a methodology for decomposing a semantic term in its perceptually distinct sub-representations, there exist certain limitations to that approach. Firstly, the process of finding sub-representations through the use of coherence metrics is limited to equalisation and cannot be transferred to other audio effects, such as compression and distortion. In order for salient sub-representations to be uncovered for other audio effects distance metrics relevant to their operation will need to be used.

In addition, as is the case with most perceptual and crowdsourced studies, there is a dependence on the resulting sub-representations on the choice of input stimuli and on the test subjects. Even though the salient sub-representations found in this study have been

verified, it is not known if these would correspond to sub-representations of the same term from different datasets. That is a problem inherit in timbre research studies.

Furthermore, this study takes into consideration two terms, in *warm* and *bright*, as having sub-representations. This is mainly due to the wide use of both terms in the sound engineering and music production disciplines and their prevalent appearance in the datasets containing semantic terms. However, other terms can potentially include sub-representations and further analysis can be performed on them. That is also something that is supported by the confidence ratings presented in Section 4.4.1, which suggest that different definitions of the same term may exist.

Chapter 6

Low-dimensional semantic equalisation

6.1 Introduction

As explained in Chapter 2, two of the main problems regarding semantic definitions of equalisation parameters are the existence of multiple definitions attributed to the same term, which was addressed in Chapter 5, but also the relationship between terms, which was detailed and analysed in Chapter 4. It has also been established that despite the disagreements over the use of timbral adjectives, these terms can provide the basis for intelligent music interfaces that can simplify the music production process and expand the creativity of sound engineers and music producers.

Despite the sound engineering literature providing guidelines regarding the frequency bands that need to be modified for altering the timbre of a sound (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005), it is frequently suggested that the parametrisation of the EQ parameters is also dependent on the input audio (Izhaki, 2013).

For these reasons, an interface is proposed that will simplify the process of creative equalisation for achieving specific timbral adjectives, while adapting to the characteristics of the input audio. By modelling the settings of a parametric EQ comprised of five biquad filters placed in series, as described by Bristow-Johnson (2016), the user will be able to manipulate the timbre of an input audio signal through an adaptive intuitive interface. By projecting the timbral characteristics of different EQ settings, used to achieve the terms *warm* and *bright* (the choice for which was discussed in Section 5.1), users will be able

to navigate a two-dimensional slider in order to modify the sonic transformations of a new input sound. The system is further improved with a signal processing method of adapting the parameter mapping process to new input audio signals, by extracting the long-term average spectrum (LTAS), as detailed by Verfaille and Arfib (2000); Verfaille et al. (2006); Zölzer and Smith (2003), making the model input dependent.

6.2 Background

Since the introduction of intelligent equalisation systems and adaptive techniques, research has also focused on systems that make use of semantic terminology.

One of the earliest systems to perform semantic equalisation was proposed by Reed (2000), using inductive learning through nearest neighbour pattern matching. Training the systems through user input, and mapping new parameters based on the nearest neighbour technique, the researchers also take into account the context-dependent attribute of the equalisation task. The system was trained to adaptively equalize three descriptive terms, *brightness, darkness* and *smoothness*.

Also focusing on the simplification of EQ interfaces, systems have been trained to learn a listener's desired equalisation curves (Sabin and Pardo, 2008, 2009b; Sabin et al., 2011), where perceptual tests are implemented to evaluate how well an equalised sound can be described by a specific timbral adjective. These perceptual systems are implemented in a similar fashion by Heise et al. (2010). By measuring the correlation between the gains at each frequency band and the listener responses, the weightings for each frequency band were found, and could provide a mapping between descriptors and parameters used to achieve these. Through this process, it is possible to assess the subjective meaning that specific terms have for individual users. Pardo et al. (2012) presents such an interactive interface to quickly learn and adapt to a user's responses.

While Reed (2000) introduced new, simplified parameters that would replace the existing technical parameters of EQ interfaces, other researchers have introduced the concept of a semantic mapping to control the timbral effect of equalisation. This technique originates from the field of timbre analysis, with numerous studies making use of dimensionality reduction techniques to visualize the structure of timbre in a low-dimensional map (Solomon (1959); Grey (1977); McAdams et al. (1995); Zacharakis et al. (2011)).

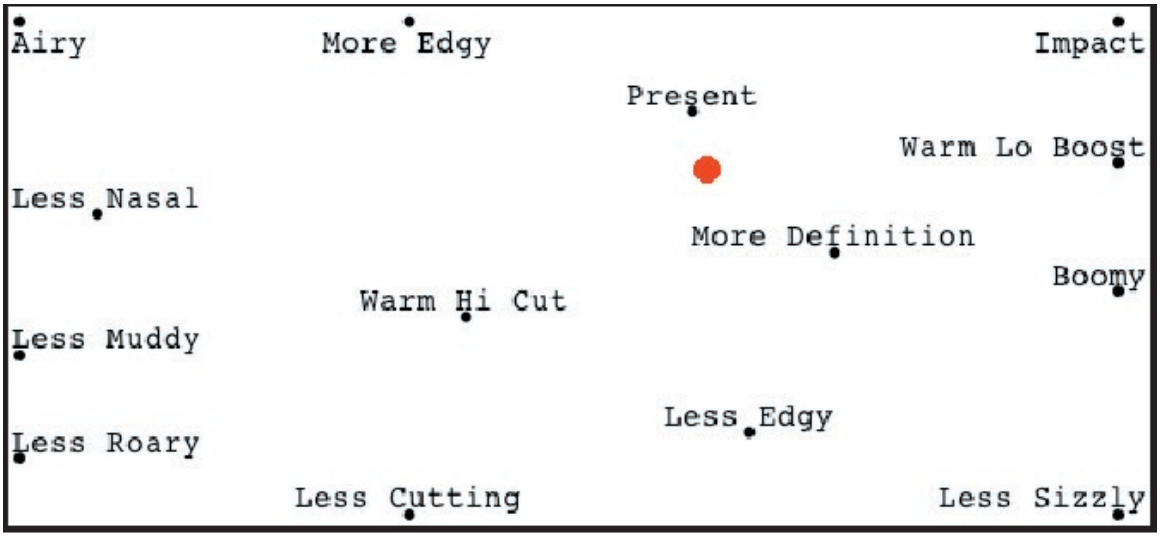
Also making use of adjectives to populate a dataset of semantic terms, Mecklenburg and Loviscach (2006) present subjEQt, a low-dimensional EQ system, using a self-organizing map architecture. The mapping was populated with descriptive terms and phrases defined by the researchers, which were considered helpful in describing the specific equalisation effect (Figure 6.1).

Combining the weighted-transfer function (Sabin and Pardo, 2008, 2009b; Sabin et al., 2011) and the notion of controlling an EQ through a semantic map (Mecklenburg and Loviscach, 2006), Sabin and Pardo (2009a) proposed 2DEQ, a simple, intuitive interface for semantically controlling an EQ (Figure 6.2). The interface can be operated through a simple two-dimensional plot, which contained four different semantic areas (*bright, tinny, warm* and *dark*). Using the weighted-transfer function information, the system was trained through user evaluation of five different instrument samples. In the case of 2DEQ, the user is able to navigate the semantic low-dimensional space by selecting an area of the interface, which sequentially controls a 40-band graphic equaliser, trained through PCA reduction and reconstruction.

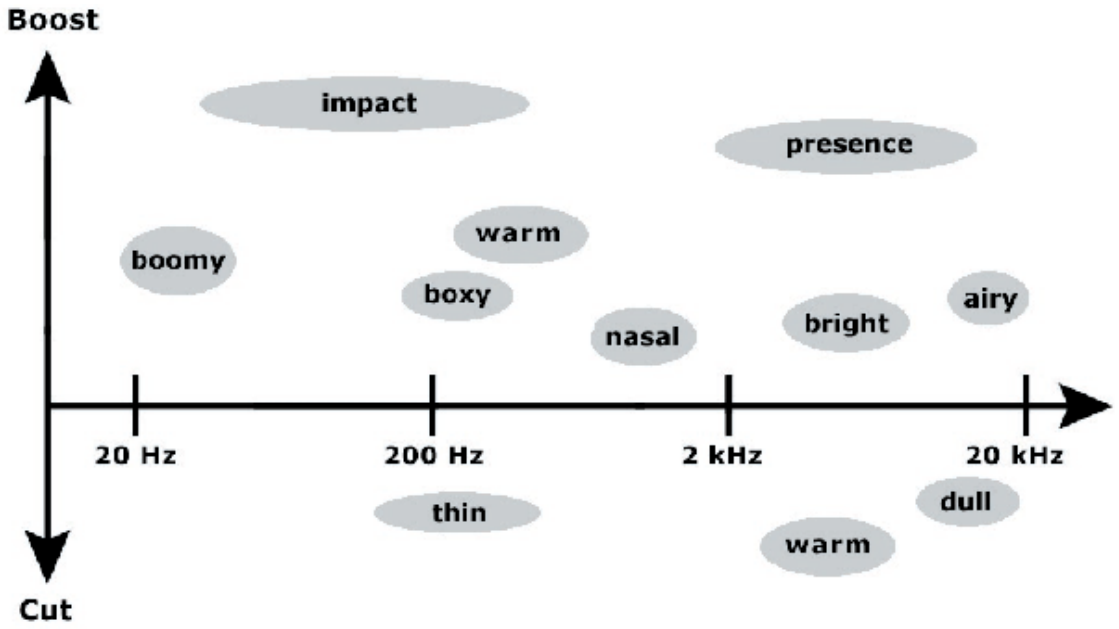
6.3 Experiment Design

In previous chapters, an in-depth analysis on different aspects of semantic processing has been performed, as well as an investigation regarding the meanings derived from timbral adjectives, and their relationship with EQ curves and parameter settings. In this chapter, a novel system is proposed that will allow for the navigation between different timbral adjectives, while adapting to new input audio information. The proposed system will replace the technical parameters of an EQ with a low-dimensional semantic space that the users can intuitively navigate.

First, data for performing semantic transformations of the two most prevalent descriptors, *warm* and *bright*, is collected. The dataset for training the model comprises 800 semantically annotated EQ parameter settings. For this task 40 participants were asked to equalise 10 musical instrument samples to achieve the two timbral adjectives. The samples featured electric guitar recordings across two genres, *metal* and *blues*. All participants were aged between 18 and 40, displayed no hearing impairments, and had at least 3 years of music production experience. The samples were presented to the subjects in a DAW



(a)



(b)

Figure 6.1: The SubjEQt interface, as was presented by Mecklenburg and Loviscach (2006), with the above figure depicting the low-dimensional space, and the below figure showing the frequency components that need to be cut/boosted to achieve each descriptive term.

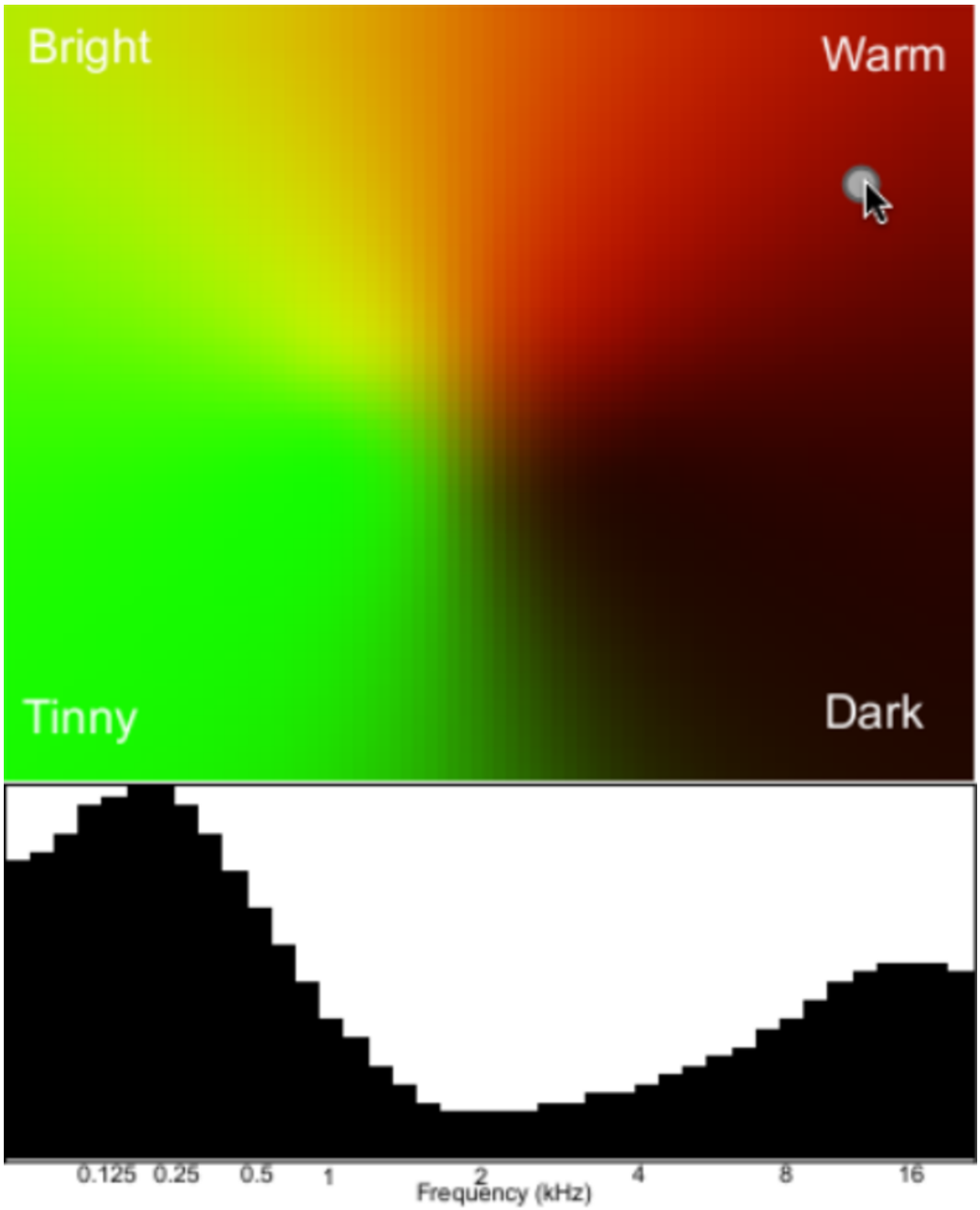


Figure 6.2: The 2DEQ interface, as was presented by Sabin and Pardo (2009a), where the low-dimensional space is presented in the above figure, and the resulting EQ curve in the below figure.

environment, where they were asked to use the SAFE EQ to achieve the two terms. After each transformation was applied, the data was recorded and the EQ settings were reset to their original positions. Additionally, to avoid inconsistencies, the musical instrument samples were all performed unaccompanied (monophonic), RMS normalisation was applied, and their range was between 20 and 30 seconds in length.

The parameters were gathered through a modified version of the SAFE data collection architecture. Figure 6.3 illustrates the architecture by which SAFE is able to capture audio feature data, before (A) and after (A') the sound has been processed by the effect ($F(x)$), as well as the interface parameters (P), which are captured and stored in an SQL database.

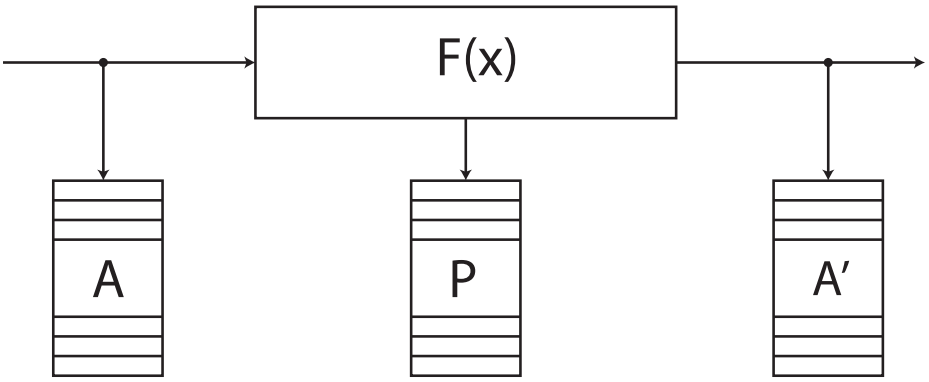


Figure 6.3: The SAFE feature extraction process.

The dimensions of the dataset are defined as the parameter settings of the EQ system. As was discussed in Section 4.3.1, the SAFE EQ is comprised of five filters arranged in series, which include one low-shelving filter (LS), one high-shelving filter (HS) and three peaking filters (P_n). In total, the dataset includes 13 dimensions, each corresponding to a different parameter of the SAFE EQ. The peaking filters can be parametrised in terms of their centre frequency, gain and Q factor, while the HS and LS filters allow for parametrisation of their centre frequency and gain, while their Q is set to 0.71. The ranges of all filters have been illustrated in Table 4.1.

In order to produce the low-dimensional interface, a process of dimensionality reduction and parameter reconstruction will be implemented. For this reason, a comparative review of dimensionality reduction and reconstruction techniques is considered, in order to find the most appropriate application for the specific dataset.

Additionally, the interface should not be agnostic to new input audio information, and

for this reason it should feature an adaptive architecture. For this to be performed relevant audio features need to be detected, measured and applied to the system. Finally, the proposed interface is presented to subjects and its performance is evaluated in achieving semantic transformations of the two terms, in *warm* and *bright*.

6.4 Model

In the context of semantic equalisation, a common trend is the simplification of the existing interfaces, mainly by reducing the original parameter space, which has been performed for the subjEQt (Mecklenburg and Loviscach, 2006) and 2DEQ (Sabin and Pardo, 2009a) projects. However, these systems focus on the relationship between different terms, without taking into account information inherent to the input audio (audio features).

For this research a model influenced by the works of Mecklenburg and Loviscach (2006) and Sabin and Pardo (2009a), is implemented that will allow for users to control a full EQ through a low-dimensional space. However, in this instance the system is trained to reconstruct parameters of two distinct timbral adjectives. As depicted in Figure 6.4, apart from the data gathering and preparation (scaling), a weighting process is applied to make the system input dependent. Following that process the system incorporates dimensionality reduction and parameter reconstruction in order to find connections between the high-dimensional parameter space and its two-dimensional representation. Finally, after the system has been trained, new user input can be passed through its functions, which will be unweighted and rescaled to account for the original transformations, and produce new high-dimensional parameters. This will allow for controlling the EQ through a timbral space (two dimensional plane) rather than the original, technical parameters (i.e. high-dimensional space).

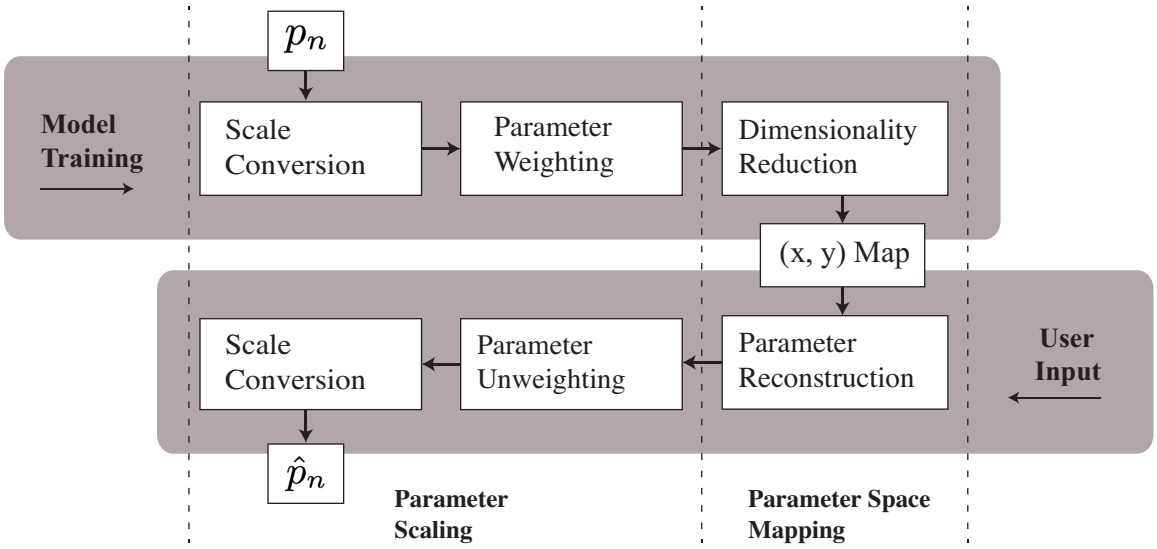


Figure 6.4: An overview of the proposed architecture for low-dimensional equalisation. The grey horizontal paths represent training and implementation phases.

In more detail, to model the relationship between the original high-dimensional parameter space and the low-dimensional mapping, a number of issues needs to be addressed. The dimensionality reduction needs to ensure a valid representation that takes into account the structure of the high-dimensional EQ parameters. This means that the low-dimensional space will preserve the underlying timbral characteristics in the original dataset, therefore allowing the user to modify an incoming signal in a musically meaningful manner.

Therefore, it should be possible to map the high-dimensional parameters by achieving a minimal reconstruction error, given a new set of (x, y) coordinates. As the loss of information is inherent in dimensionality reduction, the process of approximating the original parameters through a reconstruction stage becomes non-trivial. In order to measure the performance of the system and identify whether it is an optimal option, a comparative review is performed with a combination of 7 dimensionality reduction techniques and 6 reconstruction methods, including a model of sAE, as presented in Chapter 5.

In addition, it is essential that the process of timbral transformation through the low-dimensional plane should not be agnostic to the characteristics of the incoming audio signal. This suggests that the mapping between the low-dimensional space and the EQ parameters should be expressed as a function of both the user selected (x, y) coordinates and information extracted from the signal's spectral energy. For this reason, a series of weights

are derived based on a selection of features extracted from the LTAS coefficients of the signal. The weights are stored as a vector $a_m(A)$, computed from the input signal long-term spectral energy (A) and are then applied to the original parameter vector P, to derive the weighted parameter vector P'.

Finally, to evaluate the performance of the model through the different conditions set by this study both objective and subjective measurements are applied. These are based on the reconstruction of the input space and the structural preservation presented in the low-dimensional mapping.

Parameter scaling

Due to the different ranges that each setting possesses, if a normalisation or scaling procedure is not implemented, the algorithm can be biased towards the dimension displaying the highest variance. For this reason all the parameters/dimensions of the dataset are converted to a range of $0 < p_n < 1$. This ensures that the system will not be altered due to the existence of different ranges in the dataset. In addition, a rescaling process needs to take place before the parameter values are passed to the EQ (final step of the User Input section of Figure 6.4) in order to ensure that the parameters will be of the required range to appropriately alter the EQ characteristics.

Equation 6.1 displays the scaling process, where the p_{min} and p_{max} represent the minimum and maximum values for each parameter, while q_{min} and q_{max} represent the target range, in this case 0 and 1. During the rescaling process, these values are exchanged, so that q_{min} and q_{max} represent the minimum and maximum values, while p_{min} and p_{max} represent 0 and 1.

$$\rho_n = \frac{(p_n - p_{min})(q_{max} - q_{min})}{p_{max} - p_{min}} + q_{min} \tag{6.1}$$

Finally, a sorting algorithm is implemented to place the three mid-band filters in ascending order based on their centre frequency, so as to prevent normalisation errors due to the frequency ranges being rearranged by the users.

Parameter weighting

The parameter weighting stage is performed by selectively extracting features from the LTAS before and after the filter has been applied, which is made possible due to the SAFE architecture. The weights (a_m) gathered from the features can be considered a function of the LTAS, where their definition will vary based on the parameter representation (in this case gain, centre frequency and bandwidth) of the corresponding filter. The LTAS is therefore applied to prevent the parameters from adapting on each new frame being read. Each weighting is defined as the ratio between a spectral feature, relevant to a filter parameter, taken from the filtered audio signal (A'_k) and the signal filtered by an enclosing rectangular window (R_k). In this instance, the rectangular window is bounded by the minimum and maximum frequency values attainable by the observed filter $f_k(A)$.

The EQ can be defined as a series of biquad functions, depicted in Equation 6.2.

$$f_k = f_{k-1}(A, \vec{P}_{k-1}) \tag{6.2}$$

$$k = 1, ..., K-1 \tag{6.3}$$

where $K = 5$ represents the number of filters used by the EQ and f_k is the k_{th} biquad function, which can be defined by its transfer function, as presented in Equation 4.15.
The LTAS can then be modified by the filter in Equation 6.4 and the weighted parameter vector is retrieved by using the function displayed in Equation 6.5.

$$A'_k = \left|H_k(e^{jw})\right| A_k \tag{6.4}$$

$$p'_n = a_m(k)p_n \tag{6.5}$$

where the p_n is the n^{th} parameter in the parameter vector P. The weighting function a is then indexed by the variable m, where $m = 0$ represents the gain parameter, $m = 1$ represents centre frequency and $m = 2$ represents the Q factor. The weights for the gain parameters are calculated from the ratio of the spectral energy in the filtered spectrum (A') to the spectral energy in the rectangular window (R_k), displayed in Equation 6.6 and the process is illustrated in Figure 6.5. Finally, i is the index that iterates over the audio frames.

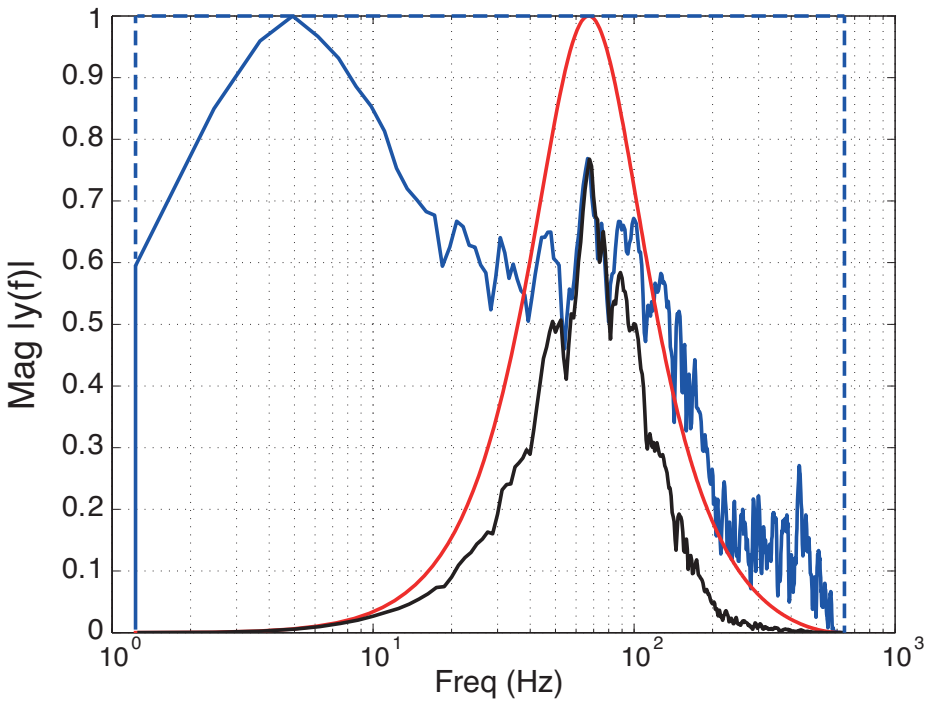


Figure 6.5: The spectrum taken from an input example signal, weighted by the biquad coefficients. The red line represents a peaking filter, the black line represents the biquad-filtered spectrum, and the blue line represents the spectral energy in the rectangular window (R_k).

$$a_0(k) = \frac{\sum_i (A'_k)_i}{\sum_i (R_k)_i} \tag{6.6}$$

For the centre frequency parameters ($m = 1$) the weights are derived as the ratio of the spectral centroids of A' and R_k respectively, where bn_i are the corresponding bins (Equation 6.7).

$$a_1(k) = \left(\frac{\sum_i (A'_k)_i bn_i}{\sum_i (A'_k)_i} \right) / \left(\frac{\sum_i (R_k)_i bn_i}{\sum_i (R_k)_i} \right) \tag{6.7}$$

The weights for the bandwidth (Q) parameters ($m = 2$) are defined as the spectral spread ratio displayed by A' and R_k. Equation 6.8 displays this function, where $(x)_{sc}$ represents the spectral centroid of x.

$$a_2(k) = \left(\frac{\sum_i \left(bn_i - (A'_k)_{sc}\right)^2 (A'_k)_i}{\sum_i (A'_k)_i} \right) / \left(\frac{\sum_i \left(bn_i - (R_k)_{sc}\right)^2 (R_k)_i}{\sum_i (R_k)_i} \right) \tag{6.8}$$

Finally, the unweigthed parameters can be retrieved during the implementation stage, by simply multiplying the weighted parameters with the inverse weights vector (Equation 6.9), where $\hat{p}$ the reconstructed approximation of p, after the dimensionality reduction has been implemented.

$$\hat{p}_n = a_m^{-1}(k) p'_n \tag{6.9}$$

Dimensionality reduction methods

Following the weighting of the filters by the features of the audio signal, the EQ parameters can be mapped into a two-dimensional space, by using dimensionality reduction. In order to find the most suitable technique for the reduction of parameters, six algorithms are comparatively evaluated on their performance. The techniques implemented in this study are part of the dimensionality reduction toolbox, developed by Van der Maaten (2007).

For dimensionality reduction, the following techniques were used:

- PCA, a very common method of embedding data into a linear subspace of reduced dimensionality, by computing the eigenvectors of the covariance matrix and maximising the variance in the mapping, originally presented by Hotelling (1933).

- Kernel PCA (kPCA), a non-linear manifold mapping technique, variation of the original PCA method, where the eigenvectors are computed from a kernel matrix instead of the covariance matrix, proposed by Schölkopf et al. (1998). In this computation of kPCA, the eigenvectors are computed from a Gaussian kernel function.

- probabilistic PCA (pPCA), another variation of PCA, which considers the standard PCA as a latent variable model. This method implements an Expectation Maximisation (EM) algorithm, a method for finding the maximum likelihood estimate of the parameters in an underlying distribution of a dataset, dependent on unobserved latent variables (Bilmes et al., 1998), suggested by Roweis (1998).

- Factor Analysis (FA), a statistical analysis that detects the relationship between different variable parameters in the dataset and groups them by the correlation of existing underlying factors (Khosla, 2004).

- Diffusion Maps (DM), a technique influenced by the study of dynamical systems, reducing the number of dimensions by embedding the original dataset in a low-dimensional mapping by retrieving the eigenvectors of Markov random walks (Nadler et al., 2006).

- Linear Discriminant Analysis (LDA), a rare example of a supervised algorithm for dimensionality reduction, where the high-dimensional dataset is projected to a linear subspace where the separability between data points belonging to different classes is maximal (Fisher, 1936).

All techniques apart from LDA have no limitations as to the number of dimensions they are reducing a dataset to. However, an inherent intricacy of LDA is that the dimensions of the subspace is limited to a value of $C - 1$, where C is the number of classes. This is performed because otherwise the technique would assign each class to a different dimension in order to achieve perfect separability. In this instance the second dimension needs to be constructed arbitrarily for this two-class dataset. All the other algorithms reduce the original dataset to a mapping comprised of two components (dimensions).

Parameter reconstruction methods

Following the projection of the high-dimensional parameters to a low-dimensional mapping, the process of approximating the original parameters is performed. Since the model will be used as the basis for a novel interface, a function that will transform the low-dimensional coordinates into high-dimensional parameters needs to be implemented. In this manner any new coordinates passed by the user will be upscaled and control the EQ.

For the reconstruction process the following techniques are tested:

- Linear Regression (LR), a technique in which a linear function is used to estimate latent variables. This method is known for its ability to model linear relationships between input and output variables.

- Natural Neighbour Interpolation (NaNI), a technique implementing Voronoi tessellation for interpolating between scattered data points, originally proposed by Sibson (1981). The method was used by Mecklenburg and Loviscach (2006) for a similar, low-dimensional EQ application.

- Nearest Neighbour Interpolation (NeNI), an interpolation technique where the new query point assumes the value of the nearest neighbour (Olivier and Hanqiang, 2012).

- Linear Interpolation (LI), an interpolation method that assumes a linear relationship between the instances in a dataset.

- Support Vector Regression (SVR), a non-linear kernel-based regression algorithm (Drucker et al., 1997), for which a Gaussian kernel function is chosen.

Stacked autoencoders

A technique that can implement both a reduction of parameters and a reconstruction of the original space is a model of sAE, which was introduced in Chapter 5. As illustrated in Figure 5.4, the model offers the versatility of isolating its decoder stage (responsible for the reconstruction), which can be then used for the implementation process.

The autoencoder for this application was built using the Theano Python library (Bergstra et al., 2010), achieving an error of 0.086 using a single hidden layer with $N = 2$ units. To further minimise the error a mirrored architecture [13 - 9 - 2] was empirically implemented, resulting in a lower error measurement of 0.08.

Vincent et al. (2010) argues that the accuracy of the reconstruction can be further improved when Gaussian noise is introduced to each stage of the network. Vincent et al. (2008) first proposed a stochastic version of the autoencoder model, where the input dataset was corrupted, but the uncorrupted version was still used as the reconstruction target. The training criterion (Equation 5.1) is then altered as in Equation 6.10.

$$RE = -logP(x|c(\widetilde{x})) \tag{6.10}$$

where x is the uncorrupted input, $\widetilde{x}$ the stochastically corrupted input, and $c(\widetilde{x})$ the representation obtained from $\widetilde{x}$. The denoising version of the sAE has been found to outperform

the standard sAE model, and perform similarly or superiorly to Deep Belief Networks (Vincent et al., 2008).

In this implementation the first unit was corrupted with Gaussian noise of 0.6 magnitude, and the second with 0.5, resulting in a further reduction of the reconstruction error to a value of 0.0784. For the optimisation of the system the RMSprop method (Tieleman and Hinton, 2012), shown to perform in a superior fashion to the standard stochastic gradient descent method (Dauphin et al., 2015), is implemented with a batch size of 10 for both the pre-training and fine-tuning parts of optimisation, and with a learning rate of 0.01 and 0.001, respectively. For the weighted parameters, it is found that a three-layer sAE model with an architecture of [13 - 9 - 6 - 2] and noise of magnitude (0.5, 0.4, 0.3) is able to outperform the two-layer denoising sAE model achieving a mean absolute error 0.058.

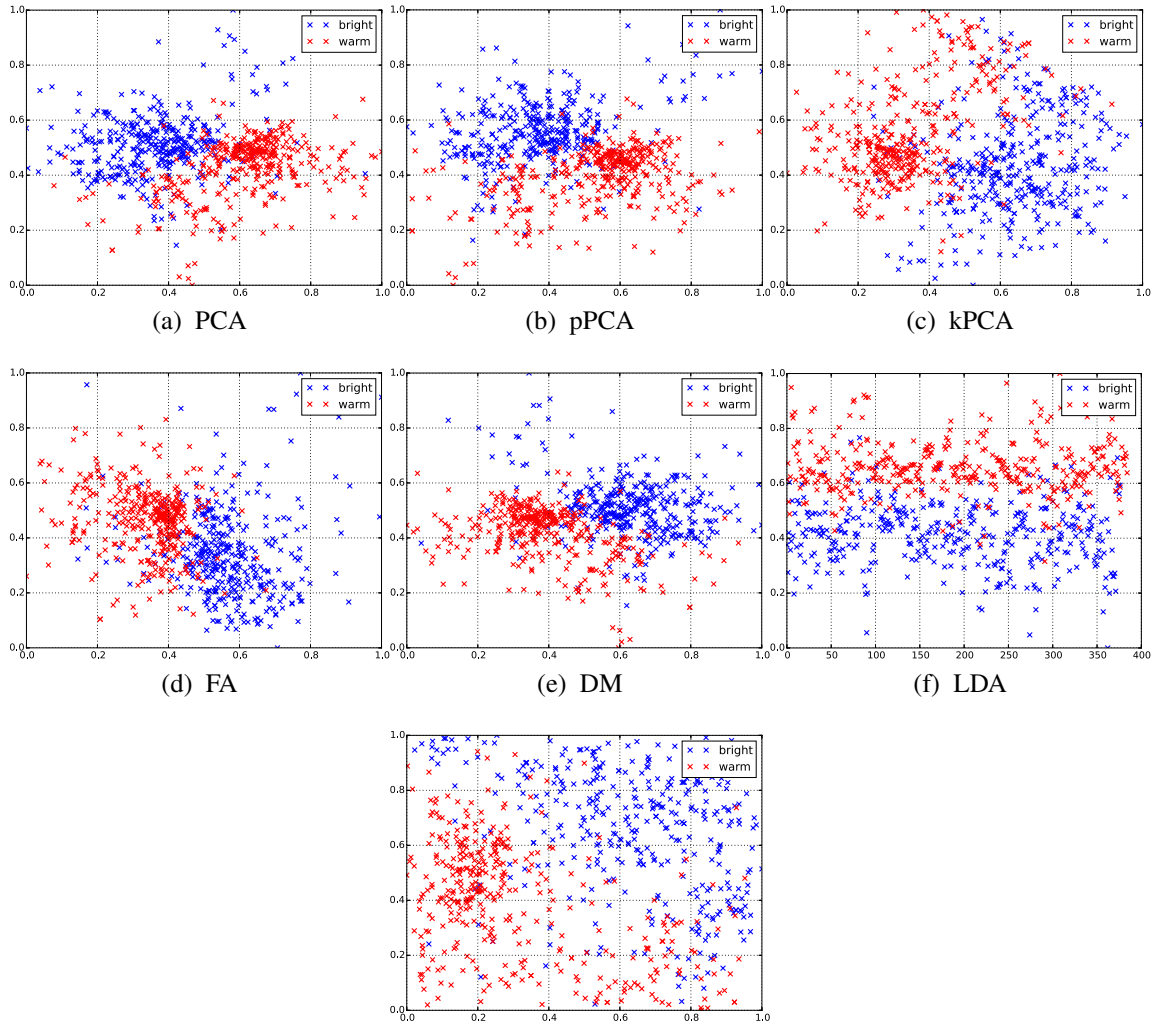
6.5 Data Preparation

The proposed system encapsulates two main processes, dimensionality reduction and parameter reconstruction. To choose the most appropriate techniques, a comparative study will be implemented to measure the system's performance. In the context of dimensionality reduction, techniques will be evaluated using the extent to which (1) the algorithm retains the structure of the high-dimensional data (trustworthiness, continuity, K-Nearest Neighbours (K-NN)) and (2) the original classes are separable in the low-dimensional mapping (Jeffries-Matusita Distance). The resulting mappings for the different dimensionality reduction techniques are presented in Figure 6.6.

Regarding the reconstruction process, the system will be evaluated as to the accuracy in the reconstruction of the high-dimensional dataset (reconstruction error).

6.5.1 Trustworthiness and continuity

The structural preservation of the different dimensionality reduction techniques can be measured through the trustworthiness and continuity metrics (Venna and Kaski, 2005). In this context, the distance of point i in the high-dimensional space is measured against its k closest neighbours in rank order, and in turn the extent to which each rank changes in the low-dimensional mapping is measured. For n samples, let $r(i, j)$ be the rank in distance of



(a) PCA (b) pPCA (c) kPCA

(d) FA (e) DM (f) LDA

(g) sAE

Figure 6.6: Two-dimensional parameter-space representations using seven dimensionality reduction techniques. The blue data points are taken from parameter spaces described as *bright* and the red points are described as *warm*.

.

sample i to sample j in the high-dimensional space U_i^k. Similarly, let $\hat{r}(i,j)$ be the rank distance between sample i and j in the low-dimensional mapping V_i^k. The resulting map is considered trustworthy if the k neighbours in the high-dimensional space are also placed close to point i in the low-dimensional mapping, as is displayed in Equation 6.11.

$$T_{(k)} = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_i^{(k)}} (r(i,j) - k) \tag{6.11}$$

On the other hand, the continuity, also known as cluster preservation, measures the extent to which the original clusters of datapoints are preserved, and can be considered as an inverse metric to trustworthiness, detecting sample points that are close to point i in the low-dimensional mapping, but not in the high-dimensional space, as shown in Equation 6.12.

$$C_{(k)} = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in V_i^{(k)}} (\hat{r}(i,j) - k) \tag{6.12}$$

Both of these equations make use of a normalising factor to bound the scores of trustworthiness and continuity between 0 and 1. The two metrics evaluate the extent to which the structure of the original dataset is preserved in the low-dimensional map over a number of k neighbours. As Van Der Maaten et al. (2009) suggests, a successful mapping of high-dimensional datapoints is achieved when the local structure of the original dataset is preserved. That suggests that the global structure is not as important as the local structure when assessing the performance of dimensionality reduction techniques.

Unweighted Model

Table 6.1 shows that the technique that achieves the highest trustworthiness rating (0.8426) is pPCA, with the sAE also performing to a similar level with 0.842. The rest of the techniques are also able to achieve high scores, ranging from 0.81 for kPCA to 0.839 for the standard form of PCA. The only technique that performs significantly worse is LDA. This is an expected result, as LDA will attempt to maximise the separation of the classes in the low-dimensional mapping instead of preserving the structure of the original dataset. For continuity the majority of the techniques perform at a similar level, with scores ranging

from 0.943 for the sAE to 0.958 for kPCA. The behaviour of the LDA for the continuity is similar to its inferior performance in the trustworthiness metric, achieving a score of 0.868, due to its reduction process.

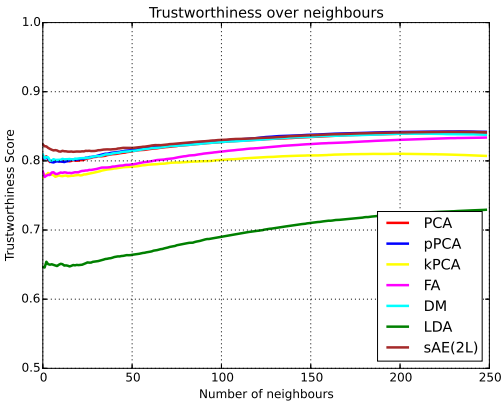
Technique	Trustworthiness	Continuity	*1*-NN Classification
Original	——	——	**91.21%**
PCA	0.8398	0.9541	87.61%
pPCA	**0.8426**	0.9567	87.92%
kPCA	0.8102	**0.9583**	86.14%
FA	0.8337	0.9490	86.19%
DM	0.8395	0.9533	87.89%
LDA	0.7292	0.8684	85.40%
sAE	0.8420	0.9439	84.01%

Table 6.1: Trustworthiness and continuity scores and *1*-NN accuracy for the different unweighted dimensionality reduction techniques.

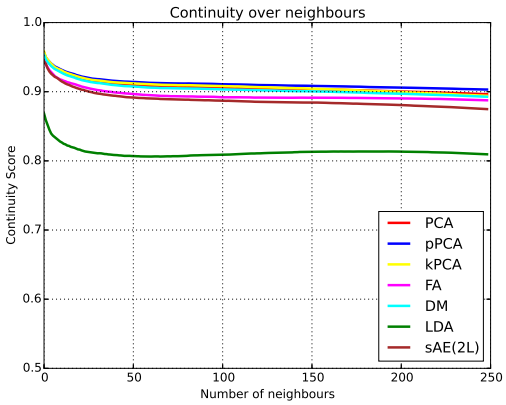
A feature of the trustworthiness and continuity metrics is their calculation over a varying number of neighbours. In this instance, the two metrics were taken from a range of 1 to 250 neighbours, in order to analyse the extent to which the technique preserves the local structure (Figures 6.7 and 6.8). The sAE in this instance achieves the highest trustworthiness scores for a lower number of neighbours (< 120), as displayed in Figure 6.7a, suggesting that the system is better at retaining the local structure of the data, a necessary attribute for a successful mapping technique. Moreover, while the continuity score of the autoencoder is lower than the remaining dimensionality reduction techniques (Table 6.1), its performance is worse by only 0.015 than the best performing technique in terms of continuity (kPCA), which can be deemed negligible.

Weighted Model

Additionally, the weights applied to the parameters improve the trustworthiness of the low-dimensional mapping when using PCA, pPCA, kPCA, DM and sAE, whilst FA and LDA exhibit significantly lower scores (Table 6.2). However, the continuity of the systems displayed minor changes, with pPCA, kPCA, DM, FA and sAE showing slight reductions, LDA showing a significant reduction in accuracy, while PCA was the only technique to exhibit an improvement. The sAE still appears more trustworthy than the other techniques for

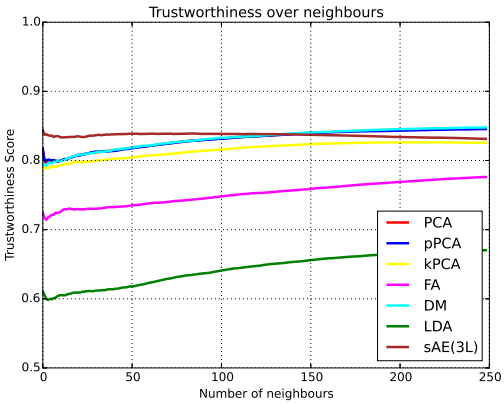


(a)

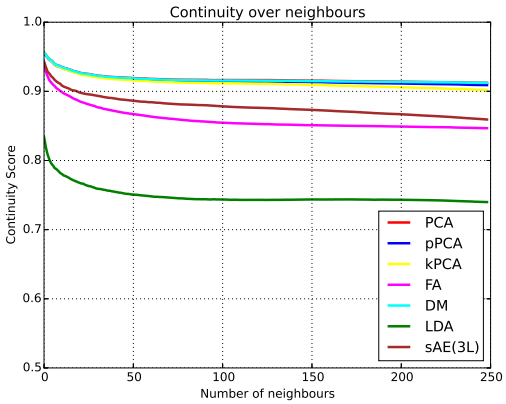


(b)

Figure 6.7: Trustworthiness and continuity plots across the different unweighted dimensionality reduction techniques.



(a)



(b)

Figure 6.8: Trustworthiness and continuity plots across the different weighted dimensionality reduction techniques.

a lower number of neighbours (Figure 6.8a), while in terms of the continuity the sAE performs better than FA and LDA, but less favourably compared to the remaining techniques (Figure 6.8b).

Technique	Trustworthiness	Continuity	*1*-NN Classification
Original	——	——	**84.9%**
PCA	0.8463	0.9562	67.85%
pPCA	0.8454	0.9552	67.39%
kPCA	0.8263	**0.9566**	69.40%
FA	0.7761	0.9359	59.52%
DM	**0.8477**	0.9561	66.03%
LDA	0.6702	0.8340	73.92%
sAE(3-Layer)	0.8440	0.9431	73.51%

Table 6.2: Trustworthiness and continuity scores and *1*-NN accuracy for the different weighted dimensionality reduction techniques.

6.5.2 K-NN

To measure the similarities between the inter-class structure of the high-dimensional space and low-dimensional mapping, Sanguinetti (2008) suggests applying a K-NN classifier with $k = 1$, and measuring the differences in classification accuracy. In this review, the nearest neighbours are found using a Euclidean distance metric with 13 and 2 dimensions, for the high-dimensional and low-dimensional spaces respectively. A K-fold cross validation with $k = 20$ is performed to derive the accuracies, with 20% of the data partitioned for testing. The technique is performed as a supervised structure preservation metric, measuring the extent to which the between-class structure is preserved in the dimensionality reduction process.

Unweighted Model

The original dataset achieves an accuracy of 91.21% in classification of 1-NN for 100 iterations of the algorithm. However, none of the dimensionality reduction techniques are able to recover this accuracy, with pPCA achieving the highest score (87.92%), as is displayed in Table 6.1. The sAE in this instance achieves the lowest accuracy (84.01%), with an additional 7.2% error from the classification algorithm in the high-dimensional dataset.

This result suggests that the sAE is not as capable as the other techniques in preserving the relationship between classes in the low-dimensional mapping.

Weighted Model

For the weighted model, the 1-NN measurements displays a drop in accuracy, with the original dataset achieving an accuracy of 84.9%. Most of the dimensionality reduction techniques are unable to preserve the classes in the low-dimensional space, achieving values as low as 59.52% (FA) to 67.85% (PCA). The only techniques that are able to surpass the 70% mark are LDA (73.92 %), something that is not surprising since the technique is built to maximise separability of classes, and the 3-Layer sAE, which scores an accuracy of 73.51%.

6.5.3 Jeffries-Matusita distance

To evaluate the extent to which the descriptors exist in opposing areas of the the low-dimensional space, a distance metric measuring the extent of separability between classes can be implemented. For measuring separability of classes Theodoridis et al. (2010) suggest the Bhattacharya distance (BH, Equation 6.13), initially proposed by Bhattacharyya (1943). However, Bruzzone et al. (1995) instead suggest the Jeffries-Matusita distance (JMD) as a more appropriate alternative, as it bounds the scores of the BH on a range of 0 (no separability) to 2 (perfect separability), displayed in Equation 6.14.

$$BH_{i,j} = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{S_i + S_j}{2} \right)^{(-1)} (\mu_i - \mu_j) + 0.5ln \left(\frac{0.5(|S_i + S_j|)}{\sqrt{|S_i|\,|S_j|}} \right) \tag{6.13}$$

$$JMD_{i,j} = \sqrt{2(1 - e^{-BH_{i,j}})} \tag{6.14}$$

where μ the mean and S the covariance matrix of classes i and j.

While class-separability is not necessarily correlated with accurate preservation of structure, a high separability score will allow users to effectively modulate between the two timbral descriptors.

Unweighted Model

In measuring the class separation through the JMD metric (Equation 6.14) for the unweighted model, it is found that kPCA outperforms the remainder of the techniques implemented, achieving a score of 0.607, while the optimised sAE model performs slightly less favourably with a score of 0.558 (Table 6.3). The remaining techniques achieve scores ranging from 0.5152 (pPCA) to 0.4862 (FA). LDA was excluded from the class separation measurement for two reasons: (1) it is a supervised technique that specifically maximises the separability between the different classes in the low-dimensional mapping, and (2) in the context of this study, LDA has reduced the dataset to a single component (dimension), while all the other techniques have reduced the dimensionality to two components. Here, there is a slight decline in the separation between classes, similar to the one identified in the 1-NN classification.

Separability Measure	**PCA**	**pPCA**	**kPCA**	**FA**	**DM**	**sAE**
JMD	0.5142	0.5152	**0.6076**	0.4862	0.5125	0.5581

Table 6.3: Jeffries–Matusita Distance (JMD) scores showing separation across different dimensionality reduction techniques for the unweighted model.

Weighted Model

The JMD metric for the weighted parameter spaces, suggests that kPCA is able to perform superiorly to the remaining dimensionality reduction techniques, achieving a score of 0.5093 (Table 6.4). The sAE model performs less favourably, achieving a score of 0.4849, while PCA, pPCA and Diffusion Maps perform to a similar level, ranging from 0.4787 to 0.4428.

Separability Measure	**PCA**	**pPCA**	**kPCA**	**FA**	**DM**	**sAE**
JMD	0.4787	0.4428	**0.5093**	0.4734	0.4702	0.4849

Table 6.4: Jeffries–Matusita Distance (JMD) scores showing separation across different dimensionality reduction techniques for the weighted model.

6.5.4 Reconstruction error

The reconstruction accuracy is the process of attaining an approximation of the original dataset from the low-dimensional mapping. This is measured through the input to output error for each pairwise combination of dimensionality reduction and reconstruction techniques. For this purpose, the mean absolute error metric is computed between the predicted (approximated, $\hat{x}_i$) and actual (original, x_i) parameter values, over the number of n samples (Equation 6.15). A K-fold cross validation with $k = 20$ iterations is implemented, with a test partition of 20% (160 training examples).

$$MAE = \frac{\sum_{i=1}^{n} |\hat{x}_i - x_i|}{n} \tag{6.15}$$

Unweighted model

The sAE is able to outperform all the other combinations of techniques, as can be seen in Table 6.5 achieving an overall error of 0.074. The second best technique in the reconstruction process has been the Diffusion Maps in combination with support vector regression, achieving an error of 0.080. The sAE is able to reconstruct the most parameters of the EQ (6) more accurately than any other combinations of techniques. The Diffusion Maps with support vector regression scored the highest in three parameter reconstructions, while FA with support vector regression in two.

P:	LS **G**	LS **F**	P_0 **G**	P_0 **F**	P_0 **Q**	P_1 **G**	P_1 **F**	P_1 **Q**	P_2 **G**	P_2 **F**	P_0 **Q**	HS **G**	HS **F**	**Mean** (μ)
PCA-LR	0.099	0.070	0.142	0.047	0.041	0.139	0.079	0.028	0.124	0.090	0.029	0.102	0.109	0.084
LDA-LR	0.194	0.070	0.150	0.047	0.041	0.171	0.082	0.028	0.116	0.090	0.030	0.123	0.106	0.096
kPCA-LR	0.081	0.070	0.136	0.047	0.040	0.150	0.082	0.027	0.130	0.084	0.029	0.120	0.107	0.085
pPCA-LR	0.099	0.069	0.138	0.046	0.039	0.142	0.078	0.027	0.126	0.092	0.030	0.104	0.108	0.084
DM-LR	0.104	0.070	0.138	0.047	0.040	0.139	0.081	0.027	0.126	0.091	0.031	0.102	0.106	0.085
FA-LR	0.151	0.068	0.156	0.042	0.040	0.143	0.068	0.029	0.144	0.084	0.030	0.103	0.094	0.089
PCA-SVR	0.086	0.064	0.123	0.046	0.040	0.137	0.079	0.028	0.125	0.089	0.031	0.097	0.095	0.080
LDA-SVR	0.196	0.068	0.152	0.048	0.040	0.171	0.081	0.028	0.116	0.087	0.031	0.123	0.105	0.096
kPCA-SVR	0.077	0.069	0.136	0.045	0.039	0.144	0.079	0.026	0.130	0.088	0.032	0.111	0.099	0.083
pPCA-SVR	0.089	0.066	0.128	0.047	0.040	0.136	0.077	0.027	0.128	0.088	0.031	0.096	0.097	0.081
DM-SVR	0.088	0.067	0.121	0.047	0.040	0.133	0.078	0.026	0.124	0.089	0.031	0.096	0.095	0.080
FA-SVR	0.144	0.062	0.137	0.041	0.039	0.144	0.066	0.026	0.144	0.085	0.030	0.098	0.082	0.084
PCA-NaNI	0.091	0.080	0.137	0.054	0.045	0.149	0.092	0.029	0.144	0.107	0.032	0.104	0.107	0.090
LDA-NaNI	0.263	0.098	0.209	0.071	0.046	0.216	0.117	0.031	0.149	0.124	0.033	0.158	0.128	0.126
kPCA-NaNI	0.083	0.082	0.159	0.056	0.042	0.154	0.095	0.029	0.160	0.116	0.033	0.125	0.108	0.096
pPCA-NaNI	0.092	0.078	0.139	0.050	0.041	0.148	0.090	0.028	0.139	0.106	0.034	0.105	0.106	0.089
DM-NaNI	0.094	0.080	0.139	0.052	0.043	0.146	0.091	0.026	0.143	0.107	0.030	0.107	0.103	0.089
FA-NaNI	0.152	0.070	0.157	0.046	0.041	0.164	0.075	0.028	0.159	0.098	0.033	0.102	0.087	0.093
PCA-NeNI	0.099	0.093	0.163	0.060	0.047	0.177	0.106	0.030	0.162	0.123	0.035	0.121	0.121	0.103
LDA-NeNI	0.252	0.100	0.194	0.060	0.042	0.217	0.109	0.031	0.151	0.120	0.037	0.158	0.115	0.122
kPCA-NeNI	0.092	0.096	0.187	0.060	0.042	0.175	0.110	0.025	0.180	0.128	0.029	0.135	0.124	0.106
pPCA-NeNI	0.103	0.088	0.162	0.059	0.042	0.170	0.107	0.027	0.160	0.123	0.034	0.120	0.117	0.101
DM-NeNI	0.110	0.090	0.161	0.059	0.046	0.175	0.101	0.025	0.159	0.124	0.034	0.122	0.116	0.102
FA-NeNI	0.176	0.082	0.171	0.054	0.041	0.193	0.087	0.028	0.205	0.114	0.034	0.138	0.096	0.109
PCA-LI	0.092	0.078	0.141	0.055	0.042	0.149	0.095	0.026	0.143	0.114	0.033	0.108	0.108	0.091
LDA-LI	0.254	0.097	0.195	0.062	0.043	0.209	0.107	0.032	0.153	0.115	0.037	0.155	0.113	0.121
kPCA-LI	0.083	0.082	0.159	0.058	0.039	0.159	0.102	0.028	0.160	0.114	0.030	0.127	0.115	0.096
pPCA-LI	0.091	0.080	0.138	0.053	0.047	0.148	0.095	0.029	0.146	0.112	0.034	0.108	0.107	0.091
DM-LI	0.098	0.076	0.142	0.051	0.045	0.149	0.089	0.030	0.146	0.112	0.033	0.108	0.105	0.091
FA-LI	0.160	0.070	0.153	0.046	0.041	0.172	0.078	0.028	0.176	0.102	0.032	0.119	0.087	0.097
sAE(2-Layer)	0.073	0.046	0.126	0.039	0.027	0.149	0.067	0.014	0.123	0.091	0.017	0.099	0.096	0.074

Table 6.5: Mean reconstruction error per parameter using combinations of dimensionality reduction and reconstruction techniques for the unweighted parameters, with the lowest reconstruction error highlighted in grey. The final column shows the mean (μ) error across all parameters, while the model with the lowest mean reconstruction error (sAE) is highlighted in green.

Weighted Model

To evaluate the efficiency of the signal specific weights, the reconstruction accuracy of each system is measured, after the weights have been applied. As can be seen in Table 6.6, all systems exhibit an overall improvement in their reconstruction accuracy. The highest performing system in this instance is the PCA with SVR pair, achieving an error of 0.059. The sAE with the same architecture (hidden layer sizes [9, 2]) is able to achieve a reconstruction error of 0.06, an improvement from the 0.0748 error displayed with unweighted parameters. However, a modified sAE with three-layer denoising topology was able to out-

perform the two-layer model and improve the reconstruction accuracy by a further 0.002, scoring as the best system with 0.058.

P:	LS **G**	LS **F**	P_0 **G**	P_0 **F**	P_0 **Q**	P_1 **G**	P_1 **F**	P_1 **Q**	P_2 **G**	P_2 **F**	P_0 **Q**	HS **G**	HS **F**	**Mean** (μ)
PCA-LR	0.052	0.059	0.062	0.040	0.023	0.114	0.075	0.018	0.107	0.088	0.020	0.034	0.106	0.061
LDA-LR	0.149	0.068	0.116	0.047	0.022	0.118	0.083	0.017	0.101	0.088	0.020	0.028	0.105	0.074
kPCA-LR	0.039	0.066	0.056	0.043	0.021	0.113	0.084	0.016	0.112	0.089	0.021	0.035	0.105	0.062
pPCA-LR	0.054	0.066	0.062	0.042	0.022	0.111	0.074	0.017	0.108	0.090	0.022	0.036	0.110	0.063
DM-LR	0.058	0.068	0.066	0.041	0.023	0.111	0.074	0.016	0.110	0.091	0.020	0.036	0.107	0.063
FA-LR	0.149	0.062	0.141	0.035	0.021	0.111	0.063	0.015	0.066	0.075	0.022	0.024	0.091	0.067
PCA-SVR	0.046	0.059	0.059	0.041	0.021	0.111	0.071	0.015	0.103	0.087	0.021	0.035	0.099	0.059
LDA-SVR	0.155	0.070	0.120	0.047	0.023	0.121	0.081	0.016	0.109	0.094	0.020	0.027	0.104	0.076
kPCA-SVR	0.036	0.068	0.052	0.044	0.023	0.111	0.080	0.016	0.106	0.090	0.022	0.035	0.108	0.061
pPCA-SVR	0.047	0.061	0.058	0.041	0.023	0.113	0.074	0.016	0.106	0.094	0.021	0.035	0.101	0.061
DM-SVR	0.050	0.063	0.060	0.042	0.024	0.110	0.074	0.016	0.103	0.089	0.020	0.035	0.100	0.060
FA-SVR	0.141	0.050	0.136	0.036	0.023	0.108	0.058	0.017	0.064	0.075	0.019	0.024	0.092	0.065
PCA-NaNI	0.048	0.066	0.064	0.047	0.026	0.127	0.081	0.019	0.116	0.096	0.024	0.038	0.111	0.066
LDA-NaNI	0.195	0.092	0.152	0.062	0.025	0.160	0.106	0.020	0.135	0.123	0.026	0.033	0.123	0.096
kPCA-NaNI	0.038	0.075	0.061	0.051	0.026	0.137	0.098	0.020	0.120	0.102	0.024	0.039	0.110	0.069
pPCA-NaNI	0.046	0.065	0.064	0.045	0.027	0.128	0.080	0.022	0.117	0.094	0.021	0.036	0.110	0.066
DM-NaNI	0.054	0.070	0.069	0.046	0.028	0.128	0.084	0.019	0.118	0.100	0.024	0.038	0.109	0.068
FA-NaNI	0.164	0.055	0.163	0.040	0.023	0.124	0.069	0.019	0.077	0.090	0.025	0.029	0.104	0.076
PCA-NeNI	0.057	0.077	0.080	0.057	0.029	0.157	0.100	0.022	0.140	0.119	0.022	0.043	0.126	0.079
LDA-NeNI	0.195	0.096	0.157	0.063	0.027	0.157	0.105	0.023	0.132	0.122	0.027	0.032	0.123	0.097
kPCA-NeNI	0.042	0.081	0.072	0.058	0.030	0.154	0.108	0.024	0.145	0.112	0.025	0.045	0.125	0.079
pPCA-NeNI	0.054	0.072	0.076	0.055	0.027	0.155	0.097	0.022	0.137	0.110	0.022	0.042	0.130	0.077
DM-NeNI	0.059	0.075	0.084	0.053	0.030	0.158	0.095	0.022	0.143	0.114	0.025	0.045	0.129	0.079
FA-NeNI	0.185	0.064	0.190	0.047	0.029	0.144	0.085	0.020	0.091	0.109	0.025	0.033	0.117	0.088
PCA-LI	0.052	0.070	0.069	0.050	0.027	0.136	0.087	0.021	0.127	0.102	0.026	0.038	0.119	0.071
LDA-LI	0.192	0.103	0.154	0.062	0.027	0.161	0.110	0.018	0.140	0.135	0.025	0.035	0.124	0.099
kPCA-LI	0.037	0.069	0.064	0.049	0.027	0.138	0.094	0.020	0.122	0.106	0.024	0.040	0.113	0.069
pPCA-LI	0.052	0.071	0.069	0.049	0.026	0.137	0.084	0.020	0.125	0.102	0.024	0.039	0.116	0.070
DM-LI	0.054	0.070	0.070	0.046	0.029	0.132	0.085	0.020	0.121	0.099	0.024	0.037	0.113	0.069
FA-LI	0.170	0.056	0.162	0.040	0.026	0.124	0.070	0.021	0.077	0.093	0.025	0.030	0.103	0.077
sAE(3-Layer)	0.065	0.053	0.081	0.040	0.021	0.106	0.075	0.015	0.077	0.081	0.017	0.028	0.096	0.058

Table 6.6: Mean reconstruction error per parameter using combinations of dimensionality reduction and reconstruction techniques for the weighted parameters, with the lowest reconstruction error highlighted in grey. The final column shows the mean (μ) error across all techniques, while the model with the lowest mean reconstruction error (sAE) is highlighted in green.

6.6 Subjective evaluation

Once the metrics defined in Section 6.5 are implemented, the appropriate model for the system, which is capable of reducing the original dataset while preserving its structure and accurately reconstructing the input parameters with minimal error, is selected. To validate the choice, subjective user testing is performed, where participants are asked to equalise a

series of audio samples using only the reduced dimensionality interface. A snapshot of the interface used for the testing is presented in Figure 6.9. The EQ controls (high-dimensional parameters) were hidden, and no indication was given regarding the underlying distribution of the two classes. This was performed in order to evaluate whether subjects were able to identify the distinct timbral regions of the two descriptors, without the use of the standard EQ parameters. At the beginning of the test the interface was set on bypass and would start processing the sound when the subjects moved the two-dimensional slider. Whenever users would proceed to the next task the interface would return to bypass mode. For the testing, the unweighted version of the model is used. This is chosen as the unweighted model will allow a direct comparison between the EQ curves gathered in the SAFE EQ dataset, and can measure the system's accuracy in generating these through the low-dimensional interface.

Figure 6.9: Snapshot of the low-dimensional EQ interface used for the subjective evaluation test. The timbral regions for *bright* and *warm* remain hidden in order to test the subjects' ability in detecting these.

During the test, samples were presented to participants in a random order across separate DAW channels. The stimuli included unaccompanied musical instrument samples, ranging from 20 to 30 seconds in length. The samples were primarily taken from electric guitars and included a variety of genres, taken from the Mixing Secrets Multitrack Audio Dataset via the Open Multitrack Testbed (De Man et al., 2016). In total 10 participants were asked to apply equalisation to achieve a *warm* or *bright* output sound for each sample. The participants were aged 18-35, had varied music production experience, ranging from 0 to 5 years, and none displayed any hearing impairment.

After the completion of the test, the class separability measured by the JMD metric can be used to provide an indication of the subjects performance (Table 6.7), and the correlation between EQ curves generated using the original 13 parameters and curves generated from the two-dimensional slider (Table 6.8).

The low-dimensional mapping is found to have increased the separability between the *warm* and *bright* classes, achieving a score of 0.8527 surpassing the original score of 0.5581. Additionally, an increase is found between the separation of the original high-dimensional classes and the opposing low-dimensional classes, with the high-dimensional *warm* examples and the low-dimensional *bright* examples achieving a separation of 0.7719, higher than the original separation between the high-dimensional classes. On the other hand, the separability between the high-dimensional parameters and the low-dimensional mapping in the same class, reveals a low separability rating (0.0846 for the *warm* instances and 0.1439 for the *bright*). This a desirable effect as it suggests that the users were able to detect the appropriate regions for the two descriptors.

Separability	**W(13-D)/ B(13-D)**	**W(2-D)/ B(2-D)**	**W(13-D)/ B(2-D)**	**B(13-D)/ W(2-D)**	**W(13-D)/ W(2-D)**	**B(13-D)/ B(2-D)**
JMD	0.5581	0.8527	0.7719	0.6988	0.0846	0.1439

Table 6.7: Jeffries-Matusita Distance (JMD) scores showing separation for data gathered from 13-dimensional parameters and a two-dimensional interface using *warm* (W) and *bright* (B) examples. Higher scores are desirable for the first four measurements, while lower scores are better for the last two columns.

The low-dimensional Euclidean distances between class centroids (Figure 6.10) and the strong positive correlation between EQ curves achieved through the original parameters and

Metric	B(13-D)/B(2-D)	W(13-D)/W(2-D)	W(13-D)/B(13-D)	W(2-D)/B(2-D)
Pearson correlation	0.9346	0.9247	-0.7594	-0.9121

Table 6.8: Pearson correlation between the EQ curves gathered from the high-dimensional (original parameters) and low-dimensional (2-D slider) parameter spaces.

the two-dimensional interfaces reinforce the capability of the system to perform semantic transformations through equalisation. Table 6.8 shows the Pearson correlation between the high-dimensional and low-dimensional class centroids, where the *bright* classes achieve a correlation of 0.9247 ($p < 0.001$) and the *warm* classes a rating of 0.9346 ($p < 0.001$). This similarity the EQ curves generated from using the full parameters (high-dimensional controls) and the EQ generated through using the low-dimensional space is illustrated in Figure 6.11a for the *warm* descriptor and Figure 6.11b for the *bright* descriptor. In addition, the negative correlation between the two classes in the high-dimensional space (-0.7594, $p < 0.001$) has been amplified, achieving a score of -0.9121 ($p < 0.001$).

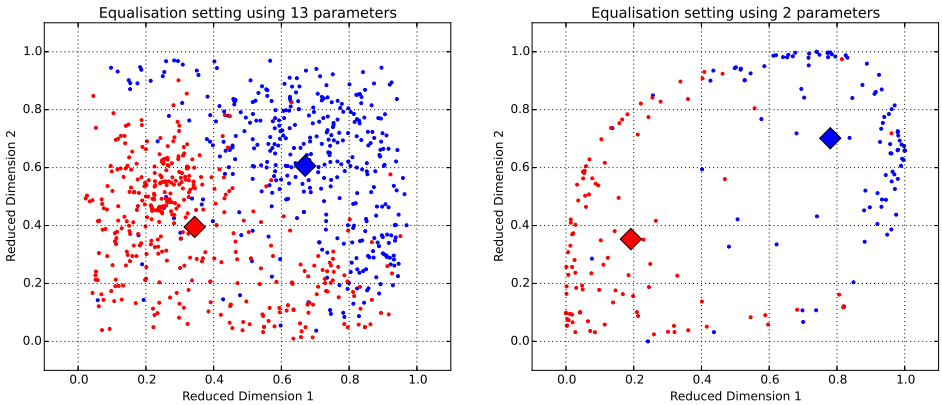


Figure 6.10: Equalisation settings for recording *warm* and *bright* samples using the original 13 parameters (left) and using the sAE-based two-dimensional EQ (right).

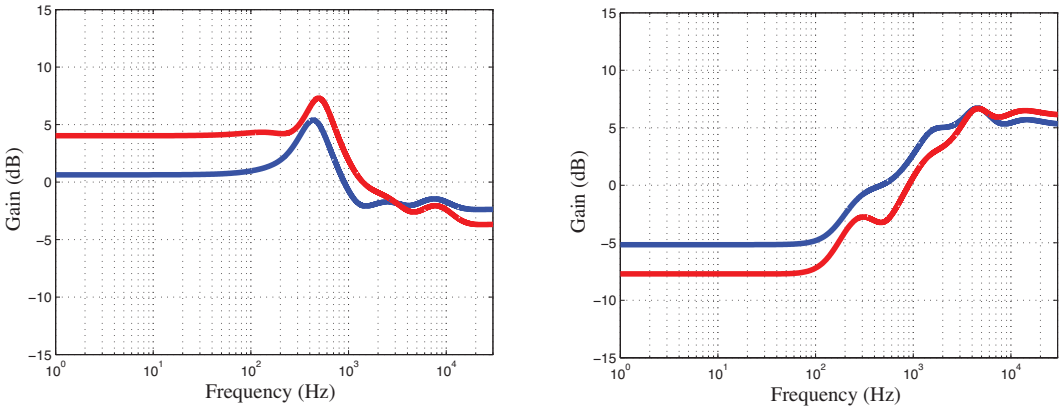
6.7 Discussion

For the reconstruction accuracy, it is found that the sAE is able to achieve better results than all other pairwise combinations of dimensionality reduction and reconstruction techniques,

whether or not the system includes the parameter weighting stage (Tables 6.5 and 6.6). In addition, the sAE achieves the second highest trustworthiness rating (Tables 6.1 and 6.2) in the low-dimensional space and also outperforms the other techniques for a smaller number of neighbours (Figures 6.7a and 6.8a), preserving the local structure. Similarly, the sAE performs to a high standard in the continuity metric, displaying a preservation of the high-dimensional clusters (Table 6.1 and Figures 6.7b and 6.8b). Furthermore, it is found that the parameter weighting reduces separability in the low-dimensional space but increases the accuracy of the reconstruction. Moreover, the system is able to reconstruct with the highest accuracy most of the parameters of the EQ (six for the unweighted and five for the weighted versions), while FA with SVR is the only other combination that is able to better reconstruct five parameters for the weighted reconstruction. However, the combination achieves lower ratings for the overall reconstruction (0.065), trustworthiness (0.7761), classification (59.52%), and a marginally lower score for continuity (0.9359).

Even though the parameter reconstruction of the sAE is sufficiently accurate for the specific application, it is bound by the intrisic dimensionality of the dataset (Kalantan, 2014), which is defined as the minimum number of components necessary to avoid significant loss of information. As was discussed in Section 5.4.1, the intrinsic dimensionality is measured using Maximum Likelihood Estimation (Levina and Bickel, 2005), which when applied to the original dataset suggests three necessary components for the *bright*/*warm* space. However, as the application requires a two-dimensional interface, the reduction procedure will be required to output two components, inherently limiting the reconstruction accuracy of the system.

Furthermore, it is found that the 2-D slider is capable of accurately modelling the EQ curves to match the timbral adjectives of the two descriptors. Table 6.8 shows that there is a very strong positive correlation between the EQ curves achieved through the high-dimensional parameter space and the EQ curves generated through navigation of the low-dimensional interface. In addition, the high-dimensional relationship between the two terms, displaying a correlation of -0.7594, is further highlighted in the low-dimensional controller, where EQ curves reach a score of -0.9121. These findings suggest that the system is able to reconstruct the salient parameters that are required to achieve both *bright* and *warm* sonic characteristics.



(a) Reconstructed *warm* EQ curve

(b) Reconstructed *bright* EQ curve

Figure 6.11: Reconstructed equaliser curves for the centroid of the (a) *warm* and (b) *bright* descriptors for both the high-dimensional (red) and low-dimensional (blue) datasets.

6.8 Conclusion

6.8.1 Overview

Regarding the simplification of the technical parameters of an EQ, a model has been presented that is capable of modulating the equalisation settings through a two-dimensional control interface. An sAE model is implemented to reduce the dimensionality of a dataset, with a weighting process to adapt the parameters to the LTAS of an input audio signal. The model is provided with semantically annotated data, and is trained to produce an approximation of its input to its output stage. New data, in the form of (x, y) coordinates in a Cartesian space can be provided by the user, with the decoder layers of the sAE producing a reconstruction relevant to the placement of the cursor in the low-dimensional mapping. The new, high-dimensional values are then unweighted and rescaled before they are passed as the new EQ parameters. The sAE model is shown to achieve a higher reconstruction accuracy than the other 30 combinations of dimensionality reduction and reconstruction (regression and interpolation) techniques, by achieving an error as low as 0.058. Additionally, the system is able to perform similarly to (and in some cases outperform) the rest of the dimensionality reduction techniques in terms of structural preservation, as illustrated through the trustworthiness and continuity measurements. Also, by subjectively testing the inter-

face, it is clear that the two-dimensional controller is able to provide an intuitive tool for the approximation of high-dimensional EQ settings extracted from the high-dimensional dataset. That is demonstrated by centroid comparison from the high-dimensional dataset and low-dimensional maps, as well as by comparing the equalisation curves in measuring the correlation.

Figure 6.12: The extended Semantic Audio Equalisation plugin with the two-dimensional interface. To modify the *brightness*/*warmth* of an audio signal, a point is positioned in two-dimensional space.

The final system is presented in Figure 6.12, where it can be implemented as an appendix to the original SAFE EQ, with the controller allowing for the swift modification of the 13-dimensional parameter space through the two timbral adjectives. In this manner, the user can interact with the 2-D interface in order to quickly alter the timbral characteristics of an input sound, or perform A/B testing between the different timbral adjectives. However, if fine-tuning is required after the 2-D interface coordinates have been set, the user can modify the high-dimensional parameters of the EQ. In this manner, the system aids the user in the creative equalisation tasks, rather than excluding them from the process.

6.8.2 Limitations

Despite the improvement over previous low-dimensional EQ systems, there are certain limitations with regard to the model presented in this chapter. Firstly, the model is limited to two terms at its current state. Even though addition of new terms might be a trivial

task, it would require that the model is retrained every time a new term is added. That is necessary to ensure that each new term will not be dependent on the model's mapping of its initial terms. In addition, in the case of expanding the terms available in the model, it is also expected that the parameter weighting stage will need to be altered in order to accommodate for audio features relevant to the new terms.

Furthermore, the parameter weighting stage posses a further level of difficulty in interpreting the results of the system, as it will force the parameter space to adapt to the features of the input audio. Moreover, when compared to the unweigthed model, the weighted parameter space displays a lower degree of separation between the different terms, suggesting that the resulting space might be more difficult to navigate.

Finally, while the parameter weighting stage presented in this chapter provides a good basis for an adaptive equalisation system, it is also limited to that audio effect alone and cannot be used for other audio effects, such as compression or distortion. The reason for this is due to the weights of the system being derived from parameters that display a relevance to the specific audio features. A similar process could still be applied to other audio effects, but audio features relevant to both the parameters of the audio effect and the semantic terminology associated with its usage would need to be found.

Chapter 7

Conclusion

This dissertation has investigated the use of semantic terminology within the context of audio effects. In order to perform this analysis a quantitative method of collecting data is considered to be optimal. By using a crowdsourcing technique a large corpus of data is gathered, which can be used for performing an analysis on the use of semantic terminology in music production practices. The subsequent analysis is performed in order to comprehend the use of audio effects in the context of timbral transformation, and to investigate the between-term and within-term relationships of semantic terminology. Finally, novel systems based on machine learning techniques are presented to facilitate in music productions tasks.

Moreover, this thesis set out to answer a series of research questions, which were presented in Section 1.3. For the first research question on whether equalisation is the primary audio effect used for altering the timbre of an input sound, this study has assessed the role of the processing module through a practical experiment. In the context of this experiment it is found that equalisation is the main tool for altering timbral characteristics in music production. In addition, the relationship of terms to audio effects has been further illustrated, and it has been found that semantic terminology is a more important factor when constructing processing chains than instrument type and genre.

For the second research question investigated in this study, regarding (a) the existence of agreement regarding the definition of semantic terms, and (b) the existence of a set underlying structure of semantic terminology, this research performed a comparative review between datasets. Through this analysis it has been found that (a) there exists an over-

all agreement regarding the definitions of timbral adjectives, and (b) there is an existing underlying structure regarding the relationships between descriptive terms. Based on this knowledge a thesaurus of timbral adjectives has been created, providing a corpus of synonymous/antonymous relationships between terms, which can be used as the basis for establishing a universal terminology of timbral adjectives. In addition, a dictionary has been constructed between descriptive terms and frequency ranges, which can aid in connecting timbral terminology with the different ranges of the frequency spectrum.

The third research question posed in this study, whether a single descriptive term can encompass perceptually divergent representations has also been addressed. By performing perceptual tests it is shown that the same descriptive term can encompass multiple discrete representations. This corresponds to the discussion regarding the use of multiple definitions as it is empirically understood by sound engineers. This finding also uncovers the complex structures natural language can take, by which the same word can be used in order to describe divergent perceptual qualities.

Finally, the fourth research question regarding the manner in which natural language and semantic terms can be used as the basis for novel audio interfaces, this study has presented two different models that can simplify the use of audio effects. A recommender system capable of producing processing chains based on a semantic term cue has been presented. Furthermore, a system that bridges the parameter settings of an EQ with perceptually relevant features of the input signal. The system is therefore able to map the parameter settings of individual descriptors, and adapt its understanding of these based on new input information. This suggests that systems can be trained to follow the actions sound engineers perform in order to alter the timbre of an input signal.

The following sections (7.1, 7.2, 7.3, 7.4) provide a more detail summary of the main findings for the analysis and implementation stages of this dissertation.

7.1 Equalisation for altering timbral characteristics

A preliminary study on processing chains was performed in order to assess the impact of individual and cascaded combinations of audio effects on predefined timbral definitions. By providing the subjects with four audio effects (i.e. equalisation, compression, distortion and reverb) and by placing no limitations with respect to the order, length and repetition

of the processing chain, it is found that the most popular processing module for altering the timbre of an input sound is equalisation. This finding appears to correlate with sound engineering and music production practices, detailed in related literature (Cousins and Hepworth-Sawyer, 2013; Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999).

In order to further analyse the effect of equalisation and its connection to natural language, the audio effect is then examined as an isolated processing unit.

7.2 Timbral definitions

A comparative review of three datasets encompassing timbral adjectives (i.e. SAFE EQ, SocialEQ and Heuristic EQ) is provided. By analysing these descriptive terms and their corresponding equalisation parameters, the confidence of term definitions between studies is assessed, in addition to comparing definitions of the same term and performing structural comparisons between term relationships. Here, it is found that confidence is not dependent on a subset of terms. While various studies assess the agreement between terms through a confidence metric, this study has shown that term confidence differs between datasets.

On the other hand, individual definitions across datasets display a positive correlation. The relationships between terms is structurally correlated between all three datasets. From this analysis, it is shown that there is an overall agreement in regard to definitions of each term. Additionally, the structural similarity between the datasets suggests that an underlying structure exists in descriptive terminology. Moreover, the chapter provides a list for term synonymity that can aid in understanding the timbral relationships in a natural language context, as well as a term-frequency relational table that can decode the meaning of different terms by considering the frequency bands they exist in.

7.3 Sub-representations

While Chapter 4 provided a study of the wide range of descriptive terms in music production, Chapter 5 considers the claim of Katz (2003) that a single timbral transformation can be achieved through a number of different approaches. To do this, the different definitions (sub-representations) that a single term can take are evaluated. By implementing a

clustering technique on the two most common descriptors in the SAFE EQ dataset, *warm* and *bright*, and then performing listening tests, it is found that subjects can accurately detect three sub-representations for each descriptor. It is shown that there is an overlap between *bright* and *warm* sub-representations, and varying degrees of correlation with the definitions that sound engineers (Izhaki, 2013; Owsinski, 2009, 2013; Katz, 2003; Cousins and Hepworth-Sawyer, 2013; Coryat, 2008; Waddell, 2013; White, 1999; Gibson, 2005) provide for each term. Finally, it is also shown that the identified sub-representations exhibit high positive correlation with other terms in the SAFE EQ dataset, which means that there exists a hierarchy of definitions, such that a term like *dark* can be considered a sub-representation of *warm*.

7.4 Perceptually informed systems

In the previous chapters, emphasis was given to the parameter settings of the EQ to achieve certain descriptive terms. In Chapter 6 a method for bridging the EQ parameters with audio features relevant to specific timbral adjectives is proposed. By implementing this model, users are able to navigate between two distinct terms, in *bright* and *warm*.

To achieve this, the sAE model has been modified to feature a denoising architecture, improving its performance. A parameter weighting stage is also considered, which allows for the relevant audio features to influence the parametrisation of the system. The model is therefore informed of both the perceptual characteristics of the new input audio, as well as the EQ parameters required to achieve timbral transformations. The system is tested through subjective evaluation, and it is found capable of achieving the necessary timbral characteristics that *warm* and *bright* descriptors exhibit.

7.5 Future research

The previous chapters of this dissertation have presented limitations regarding the research (Sections 3.7.2, 4.5.5, 5.7.2 and 6.8.2), and they also suggested ways in which these shortcomings can be bypassed. In consideration to the experience gathered from performing the analysis for this thesis, this section will focus on alternative research topics that can hopefully be examined in the future.

Firstly, while this research has performed a series of experiments in order to evaluate the findings of the individual chapters it has not investigated how these systems will operate in a studio environment and how professional sound engineers would interact with them. An interesting direction for future research would be to evaluate the findings through exposing the results regarding the definitions of semantic terms to professional sound engineers and investigate their agreement with the findings. In addition, the interfaces presented in this study could also be studied in how they would be used by professional sound engineers in a studio environment. By then gathering feedback, the interfaces can be tailored to perform in a superior fashion and meet the needs of the engineers.

In addition, this study presented a method for comparing different datasets through a series of different correlation measurements (Chapter 4). Even though the use of correlation and coherence metrics is relevant with the use of equalisation, it can also be used to compare datasets of timbral adjectives achieved through other audio effects, such as compression, distortion and reverb. Even though a different metric would need to be defined to derive the similarity between mean definitions, the relational distances and structural similarities between the datasets can be calculated in the same manner. This methodology for comparing datasets presented in Chapter 4 can be adapted even to areas outside of audio, as it provides a comparison between datasets on different levels (individual, relational and structural).

Also, this research followed a quantitative methodology for data gathering, something that other researchers have also attempted (Cartwright, 2016; Enderby, 2017). On the other hand, some researchers have taken a qualitative approach (De Man, 2017; Pestana, 2013) in order to comprehend mixing practices and in various degrees the use of semantic terms in sound engineering. While both approaches have their own advantages and disadvantages, a hybrid method has not been attempted yet. By taking advantage of systems that allow for large scale information gathering and also by using professional sound engineers, with years of experience in the discipline, these two methods can be united. This would result in a very strong database, which can act as the basis for deeply understanding the methodology sound engineers and music producers follow. The technological advances regarding the treatment of big datasets, and also the ability of web-based applications to gather information, can make such a project possible.

Furthermore, this research has displayed a way in which crowdsourced information can

be used in order to analyse the process that sound engineers follow to achieve a specific task, and also to produce novel music production tools to aid in that process. While this research was focused on individual tracks and the audio effects that are used to transform timbre, an analysis of full mixing sessions can be performed instead. The main difficulty that this project holds is the inherently high-dimensional nature of a mix sessions, which include volume controls, gains, panning, channel strips and audio effects. Still, the use of novel methods for data collection through web interfaces allow for the collection of data in a much larger scale, and might be able to bypass the problem of the high-dimensionality of mix sessions to an extent.

In addition, this research has focused on the timbral transformation of individual tracks, however mix engineers have to take into consideration the relationship between tracks in the context of a mix. While a processed track in isolation might display the desired timbral characteristics, when placed in a mix further adjustments might be required. This is a complex task that requires information to be gathered across all audio tracks in a mix, but it offers the opportunity for innovative research in cross-adaptive systems that can be used for timbral transformations.

Moreover, this study has focused on the use of timbral adjectives to creatively process an input sound. However, this is not the only creative aspect in the context of music production techniques. Each sound engineer has an individual and unique style when mixing music. Considering the novel implementations of style-transfer systems, it is possible to map these different characteristics and apply them to new input audio. This will firstly lead to uncovering what makes each different sound engineer unique and, secondly will further expand the creative potential of mixing, where combinations of individual traits can be applied.

Also, this research has focused on the creative use of audio effects, but there is still an opportunity for further research on how audio effects can be used for corrective purposes. This might not be the case for reverb and distortion effects, but equalisation and compression are widely used for correcting problems with an input sound. In the case of equalisation this is performed to remove problematic frequencies, while compression can result in a more consistent dynamic range. Some of the techniques implemented in this dissertation can be adapted for that procedure, however given that corrective tasks are much more dependent on the input audio, any research towards that direction will need to focus

heavily on adaptive architectures.

Finally, while this study has performed an analysis on the use of semantic terms in the context of music production, the use of natural language is not necessarily exclusive to timbre or just audio in general. The research can therefore be expanded to adjacent areas such as image processing, where these terms are used to describe the qualities of the processing that has been performed on an image or video file. Analysis of the terminology on that field can result in a better understanding of what those terms mean, and it can also lead to innovative tools that simplify a more complex process.

7.6 Concluding remarks

The technological advancements in previous years have made it possible to simplify the process of recording, mixing and mastering, and this trend will most certainly continue. Sound engineering and music production will also be affected by this and the industry as a whole will need to be adapted to this new reality. Not only does this mean that users lacking the technical expertise will enter these disciplines, but also that experienced producers and engineers will demand more flexibility and novelty from their equipment.

The recording industry has seen significant leaps in the past. Editing tools, such as Celemony's Melodyne, for pitch correction, and Izotope's RX series, for noise reduction and audio repair, were once considered infeasible. These innovations that belong in the area of corrective processing have been fully embraced by music producers and sound engineers, and have drastically changed production quality for the better. The same should be done for areas of creative audio. With software becoming capable of automatic mixing, it is a matter of time before it is possible to perform creative actions with the same degree of automation. Even though this might appear as if creativity itself will become automated, it should actually be considered as a way of extending the creative potential of each producer through innovative means.

The innovation of processing units can be further expanded through Web Audio implementations. Systems operating within a web-browser have been already introduced (Jillings and Stables, 2017a,b), and these can further revolutionise the practices followed in the industry. Through this process, extensive datasets of the various operations that producers and engineers perform can be generated, resulting in better understanding of an individ-

ual's actions to balance a mix, process the individual tracks through audio effects, make corrective edits, label the tracks, and so on. This provides the opportunity to simplify the complex tasks required in a mixing scenario, saving necessary time for the engineers, who would then be required to simply fine-tune the system's responses. On the other hand, for creative tasks, recommender systems can be introduced, which will enhance the creativity of the engineers, by suggesting novel methods that others have used to alter timbral characteristics, or providing them with novel interfaces that allows them to control different aspects of an input signal. In addition, these systems will allow for online cooperation, making collaboration between engineers and producers easier and faster. The applications that Web Audio can have, considering the increases in Internet bandwidth, can potentially change completely the manner in which the recording industry functions.

These changes are inevitable, and technology has reached a point where these ideas are starting to be realised. By adapting to this new reality, music production can be made less costly, and more people will be able to participate in it. As a result, this will improve the overall quality of releases, by increasing the level of creativity and minimising the time required for menial tasks.

Bibliography

Abel, J. S. and Berners, D. P. (2004). Filter design using second-order peaking and shelving sections., *Proceedings of the International Computer Music Conference*.

Alluri, V. and Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre, *Music Perception: An Interdisciplinary Journal* **27**(3): 223–242.

ANSI (1973). American national standard-psychoacoustical terminology, *Technical report*, American National Standards Institute, New York.

Barthet, M., Guillemain, P., Kronland-Martinet, R. and Ystad, S. (2010). From clarinet control to timbre perception, *Acta Acustica united with Acustica* **96**(4): 678–689.

Bauer, B. (1962). A century of microphones, *Proceedings of the IRE* **50**(5): 719–729.

Bazil, E. (2009). *Sound Equalization Tips and Tricks*, PC Publishing.

Beauchamp, J. W. (1982). Synthesis by spectral amplitude and" brightness" matching of analyzed musical instrument tones, *Journal of the Audio Engineering Society* **30**(6): 396–406.

Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering, *Advances in neural information processing systems*, pp. 585–591.

Bello, J. P. (2016). MPATE-GE 2623 Music Information Retrieval, Lectures Notes: Low-level features - Timbre Analysis, Department of Music and Performing Arts, New York University.

Bengio, Y., Lamblin, P., Popovici, D. and Larochelle, H. (2007). Greedy layer-wise training of deep networks, *Advances in neural information processing systems*, pp. 153–160.

Bengio, Y. et al. (2009). Learning deep architectures for AI, *Foundations and trends in Machine Learning* **2**(1): 1–127.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D. and Bengio, Y. (2010). Theano: A CPU and GPU math compiler in Python, *Proc. 9th Python in Science Conf*, pp. 1–7.

Berkhin, P. et al. (2006). A survey of clustering data mining techniques., *Grouping multidimensional data* **25**: 71.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distribution, *Bull. Calcutta Math. Soc* .

Bilmes, J. A. et al. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models, *International Computer Science Institute* **4**(510): 126.

Birnbaum, M. H. (2004). Human research and data collection via the internet, *Annual review of psychology* **55**.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition, *Biological cybernetics* **59**(4): 291–294.

Brandt, M. and Bitzer, J. (2012). Hum removal filters: Overview and analysis, *Audio Engineering Society Convention 132*, Audio Engineering Society.

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*, MIT press.

Brent, W. (2009). Cepstral analysis tools for percussive timbre identification, *Proceedings of the 3rd International Pure Data Convention, Sao Paulo, Brazil*, sn.

Bristow-Johnson, R. (2016). Cookbook formulae for audio eq biquad filter coefficients, *http://www. musicdsp. org/files/Audio-EQ-Cookbook. txt* .

Brookes, T. and Williams, D. (2007). Perceptually-motivated audio morphing: Brightness, *Audio Engineering Society Convention 122*, Audio Engineering Society.

Brookes, T. and Williams, D. (2010). Perceptually-motivated audio morphing: Warmth, *Audio Engineering Society Convention 128*, Audio Engineering Society.

Bruzzone, L., Roli, F. and Serpico, S. B. (1995). An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection, *IEEE Transactions on Geoscience and Remote Sensing* **33**(6): 1318–1321.

Bullock, J. (2007). Libxtract: a lightweight library for audio feature extraction., *Proceedings of the International Computer Music Conference*.

Bullock, J. (2008). *Implementing audio feature extraction in live electronic music*, PhD thesis, Birmingham City University.

Butterworth, S. (1930). On the theory of filter amplifiers, *Wireless Engineer* **7**(6): 536–541.

Caclin, A., McAdams, S., Smith, B. K. and Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones a, *The Journal of the Acoustical Society of America* **118**(1): 471–482.

Campbell, G. A. (1922). Physical theory of the electric wave-filter, *Bell Labs Technical Journal* **1**(2): 1–32.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition, *Psychometrika* **35**(3): 283–319.

Cartwright, M. (2016). *Supporting Novice Communication of Audio Concepts for Audio Production Tools*, PhD thesis, Northwestern University.

Cartwright, M. B. and Pardo, B. (2013). Social-eq: Crowdsourcing an equalization descriptor map., *ISMIR*, pp. 395–400.

Cartwright, M., Pardo, B., Mysore, G. J. and Hoffman, M. (2016). Fast and easy crowdsourced perceptual audio evaluation, *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, IEEE, pp. 619–623.

Coryat, K. (2008). *Guerrilla Home Recording: How to Get Great Sound from Any Studio (No Matter How Weird or Cheap Your Gear Is)*, MusicPro guides, Hal Leonard Corporation.

Cousins, M. and Hepworth-Sawyer, R. (2013). *Practical Mastering: A Guide to Mastering in the Modern Studio*, CRC Press.

Darke, G. (2005). Assessment of timbre using verbal attributes, *Conference on Interdisciplinary Musicology. Montreal, Quebec*.

Dauphin, Y., de Vries, H. and Bengio, Y. (2015). Equilibrated adaptive learning rates for non-convex optimization, *Advances in neural information processing systems*, pp. 1504–1512.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure, *IEEE transactions on pattern analysis and machine intelligence* (2): 224–227.

De Man, B. (2017). *Towards a better understanding of mix engineering*, PhD thesis, Queen Mary University of London.

De Man, B., Reiss, J. D. et al. (2016). The open multitrack testbed: Features, content and use cases, *Proceedings of the 2nd AES Workshop on Intelligent Music Production*, Vol. 13.

De Poli, G. and Prandoni, P. (1997). Sonological models for timbre characterization, *Journal of New Music Research* **26**(2): 170–197.

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

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

Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data, *Proceedings of the National Academy of Sciences* **100**(10): 5591–5596.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J. et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, *Ecography* **36**(1): 27–46.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J. and Vapnik, V. (1997). Support vector regression machines, *Advances in neural information processing systems*, pp. 155–161.

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

Emiroglu, S. S. (2007). *Timbre perception and object separation with normal and impaired hearing*, PhD thesis, Universität Oldenburg.

Enderby, S. (2017). *Timbral Control of Audio Through Harmonic Excitation*, PhD thesis, Birmingham City University.

Enderby, S. and Stables, R. (2017). A nonlinear method for manipulating warmth and brightness, *20th International International Conference on Digital Audio Effects (DAFx-17), Edinburgh, UK*.

Erickson, R. (1975). *Sound structure in music*, Univ of California Press.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise., *Kdd*, Vol. 96, pp. 226–231.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper, *ACM SIGKDD explorations newsletter* **4**(1): 65–75.

Ethington, R. and Punch, B. (1994). Seawave: A system for musical timbre description, *Computer Music Journal* **18**(1): 30–39.

Ferrando, P. J. and Lorenzo-Seva, U. (2013). Unrestricted item factor analysis and some relations with item response theory, *Recuperado de http://psico. fcep. urv. es/utilitats/factor/.[Links]* .

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of human genetics* **7**(2): 179–188.

Fletcher, N. H. (1999). The nonlinear physics of musical instruments, *Reports on Progress in Physics* **62**(5): 723.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis, *The computer journal* **41**(8): 578–588.

Freed, D. J. (1990). Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events, *The Journal of the Acoustical Society of America* **87**(1): 311–322.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points, *science* **315**(5814): 972–976.

Fritz, C., Blackwell, A. F., Cross, I., Woodhouse, J. and Moore, B. C. (2012). Exploring violin sound quality: Investigating english timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties, *The Journal of the Acoustical Society of America* **131**(1): 783–794.

Fuhrmann, F. (2012). *Automatic musical instrument recognition from polyphonic music audio signals*, PhD thesis, PhD thesis, Universitat Pompeu Fabra.

Gibson, D. (2005). *The art of mixing: a visual guide to recording, engineering, and production*, Cengage Learning.

Greiner, R. A. and Schoessow, M. (1983). Design aspects of graphic equalizers, *Journal of the Audio Engineering Society* **31**(6): 394–407.

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

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

Hajda, J. M. (2007). The effect of dynamic acoustical features on musical timbre, *Analysis, synthesis, and perception of musical sounds*, Springer, pp. 250–271.

Hajda, J. M., Kendall, R. A., Carterette, E. C. and Harshberger, M. L. (1997). Methodological issues in timbre research.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2002). Cluster validity methods: Part I, *ACM Sigmod Record* **31**(2): 40–45.

Heise, S., Hlatky, M. and Loviscach, J. (2010). A computer-aided audio effect setup procedure for untrained users, *Audio Engineering Society Convention 128*, Audio Engineering Society.

Herrera, P., Bello, J., Widmer, G., Sandler, M., Celma, Ò., Vignoli, F., Pampalk, E., Cano, P., Pauws, S. and Serra, X. (2005). Simac: Semantic interaction with music audio contents, *Integration of Knowledge, Semantics and Digital Media Technology, 2005. EWIMT 2005. The 2nd European Workshop on the (Ref. No. 2005/11099)*, IET, pp. 399–406.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks, *science* **313**(5786): 504–507.

Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length and helmholtz free energy, *Advances in neural information processing systems*, pp. 3–10.

Hirata, Y. (1981). Digitalization of conventional analog filters for recording use, *Journal of the Audio Engineering Society* **29**(5): 333–337.

Hood, J. L. (1997). *Valve and transistor audio amplifiers*, Elsevier.

Hopkins, B. and Skellam, J. G. (1954). A new method for determining the type of distribution of plant individuals, *Annals of Botany* **18**(2): 213–227.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components., *Journal of educational psychology* **24**(6): 417.

Howard, D., Disley, A. and Hunt, A. (2007). Towards a music synthesizer controlled by timbral adjectives, *Proceedings of 14th International Congress on Sound & Vibration*.

Howard, D. M. and Tyrrell, A. M. (1997). Psychoacoustically informed spectrography and timbre, *Organised Sound* **2**(02): 65–76.

Huber, D. M. and Runstein, R. E. (2013). *Modern recording techniques*, CRC Press.

Ilango, V., Subramanian, R. and Vasudevan, V. (2011). Cluster analysis research design model, problems, issues, challenges, trends and tools, *International Journal on Computer Science and Engineering* **3**(8): 3064.

Iverson, P. and Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbrea, *The Journal of the Acoustical Society of America* **94**(5): 2595–2603.

Izhaki, R. (2013). *Mixing audio: concepts, practices and tools*, Taylor & Francis.

Jensen, K. (1999). *Timbre models of musical sounds*, Department of Computer Science, University of Copenhagen.

Jillings, N., Bullock, J. and Stables, R. (2016). Js-xtract: A realtime audio feature extraction library for the web, *International Society for Music Information Retrieval Conference*.

Jillings, N., Moffat, D., De Man, B., Reiss, J. D. and Stables, R. (2016). Web audio evaluation tool: A framework for subjective assessment of audio.

Jillings, N. and Stables, R. (2017a). An intelligent audio workstation in the browser, *Proceedings of the Web Audio Conference, London, UK*.

Jillings, N. and Stables, R. (2017b). Investigating music production using a semantically powered digital audio workstation in the browser, *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*, Audio Engineering Society.

Jillings, N., Wang, Y., Reiss, J. D. and Stables, R. (2016). Jsap: A plugin standard for the web audio api with intelligent functionality, *Audio Engineering Society Convention 141*, Audio Engineering Society.

Jimenez, L. O. and Landgrebe, D. A. (1998). Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **28**(1): 39–54.

Kalantan, Z. (2014). *Methods for Estimation of Intrinsic Dimensionality*, PhD thesis, Durham University.

Katz, R. A. (2003). *Mastering audio: the art and the science*, Butterworth-Heinemann.

Kendall, R. A. and Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres, *Music Perception: An Interdisciplinary Journal* **8**(4): 369–404.

Kendall, R. A. and Carterette, E. C. (1993). Verbal attributes of simultaneous wind instrument timbres: Ii. adjectives induced from piston9s" orchestration", *Music Perception: An Interdisciplinary Journal* **10**(4): 469–501.

Khosla, N. (2004). Dimensionality reduction using factor analysis, *Griffith University, Australia* .

Knopoff, L. (1963). An index for the relative quality among musical instruments, *Ethnomusicology* **7**(3): 229–233.

Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering, *International Journal* **1**(6): 90–95.

Koelsch, S. (2011). Towards a neural basis of processing musical semantics, *Physics of life reviews* **8**(2): 89–105.

Krimphoff, J., McAdams, S. and Winsberg, S. (1994). Caractérisation du timbre des sons complexes. ii. analyses acoustiques et quantification psychophysique, *Le Journal de Physique IV* **4**(C5): C5–625.

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

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* **29**(1): 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method, *Psychometrika* **29**(2): 115–129.

Kulka, L. d. G. (1972). Equalization - the highest, most sustained expression of the recordist's heart, *Recording Engineer/Producer* **3**.

Kuwano, S. and Namba, S. (2001). Dimension of sound quality and their measurement, *Proceedings of the 17th International Congress on Acoustics (ICA)*.

Lafon, S. and Lee, A. B. (2006). Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization, *IEEE transactions on pattern analysis and machine intelligence* **28**(9): 1393–1403.

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

Lapointe, F.-J. and Legendre, P. (1995). Comparison tests for dendrograms: A comparative evaluation, *Journal of Classification* **12**(2): 265–282.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J. and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation, *Proceedings of the 24th international conference on Machine learning*, ACM, pp. 473–480.

Legendre, P. and Legendre, L. F. (2012). *Numerical ecology*, Vol. 24, Elsevier.

Levina, E. and Bickel, P. J. (2005). Maximum likelihood estimation of intrinsic dimension, *Advances in neural information processing systems*, pp. 777–784.

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

Licklider, J. C. (1951). Basic correlates of the auditory stimulus.

Linkwitz, S. H. (1976). Active crossover networks for noncoincident drivers, *Journal of the Audio Engineering Society* **24**(1): 2–8.

Lokki, T., Pätynen, J., Kuusinen, A., Vertanen, H. and Tervo, S. (2011). Concert hall acoustics assessment with individually elicited attributes, *The Journal of the Acoustical Society of America* **130**(2): 835–849.

Loughran, R., Walker, J., O'Neill, M. and O'Farrell, M. (2008). The use of mel-frequency cepstral coefficients in musical instrument identification., *Proceedings of the International Computer Music Conference*.

Maddams, J. A., Finn, S. and Reiss, J. D. (2012). An autonomous method for multi-track dynamic range compression, *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach, *Cancer research* **27**(2 Part 1): 209–220.

Markov, A. (1971). Extension of the limit theorems of probability theory to a sum of variables connected in a chain.

Marozeau, J. and de Cheveigné, A. (2007). The effect of fundamental frequency on the brightness dimension of timbre, *The Journal of the Acoustical Society of America* **121**(1): 383–387.

Marozeau, J., de Cheveigné, A., McAdams, S. and Winsberg, S. (2003). The dependency of timbre on fundamental frequency, *The Journal of the Acoustical Society of America* **114**(5): 2946–2957.

Martin, K. D. (1999). *Sound-source recognition: A theory and computational model*, PhD thesis, Massachusetts Institute of Technology.

Martinez, W. L., Martinez, A. R., Martinez, A. and Solka, J. (2010). *Exploratory data analysis with MATLAB*, CRC Press.

Massenburg, G. (1972). Parametric equalization, *Audio Engineering Society Convention 42*, Audio Engineering Society.

McAdams, S. (1999). Perspectives on the contribution of timbre to musical structure, *Computer Music Journal* **23**(3): 85–102.

McAdams, S., Giordano, B., Susini, P., Peeters, G. and Rioux, V. (2006). A meta-analysis of acoustic correlates of timbre dimensions, *Journal of the Acoustical Society of America* **120**(5): 3275.

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes, *Psychological research* **58**(3): 177–192.

Mecklenburg, S. and Loviscach, J. (2006). subjeqt: Controlling an equalizer through subjective terms, *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, ACM, pp. 1109–1114.

Melara, R. D. and Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness, *Attention, Perception, & Psychophysics* **48**(2): 169–178.

Miller, J. R. and Carterette, E. C. (1975). Perceptual space for musical structures, *The Journal of the Acoustical Society of America* **58**(3): 711–720.

Miranda, E. R., Correa, J. and Wright, J. (2000). Categorising complex dynamic sounds, *Organised Sound* **5**(02): 95–102.

Moffat, D., Ronan, D., Reiss, J. D. et al. (2015). An evaluation of audio feature extraction toolboxes, *18th International International Conference on Digital Audio Effects (DAFx-15), Trondheim, Norway*.

Moore, B. C., Glasberg, B. R. and Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness, *Journal of the Audio Engineering Society* **45**(4): 224–240.

Moravec, O. and Štěpánek, J. (2003). Verbal description of musical sound timbre in czech language, *organ* **4**: 3–7.

Nadler, B., Lafon, S., Coifman, R. R. and Kevrekidis, I. G. (2006). Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, *Applied and Computational Harmonic Analysis* **21**(1): 113–127.

Nykänen, A., Johansson, Ö., Lundberg, J. and Berg, J. (2009). Modelling perceptual dimensions of saxophone sounds, *Acta Acustica United with Acustica* **95**(3): 539–549.

Olivier, R. and Hanqiang, C. (2012). Nearest neighbor value interpolation, *arXiv preprint arXiv:1211.1768* .

Owsinski, B. (2009). *The recording engineer's handbook*, Nelson Education.

Owsinski, B. (2013). *The mixing engineer's handbook*, Nelson Education.

Papanikolaou, G. and Pastiadis, C. (2009). Multiple dichotomies in timbre research, *Archives of Acoustics* **34**(2): 137–155.

Pardo, B., Little, D. and Gergle, D. (2012). Towards speeding audio EQ interface building with transfer learning, *Machine Learning* **10**: 11.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11): 559–572.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N. and McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals, *The Journal of the Acoustical Society of America* **130**(5): 2902–2916.

Pestana, P. D. L. G. (2013). *Automatic mixing systems using adaptive digital audio effects*, PhD thesis, Universidade Catolica Portuguesa.

Plomp, R. (1970). Timbre as a multidimensional attribute of complex tones, *Frequency analysis and periodicity detection in hearing* pp. 397–414.

Plomp, R. (1976). Timbre of complex tones, *Aspects of tone sensation: A psychophysical study* pp. 85–110.

Plomp, R. and Steeneken, H. (1969). Effect of phase on the timbre of complex tones, *The Journal of the Acoustical Society of America* **46**(2B): 409–421.

Podder, P., Hasan, M. M., Islam, M. R. and Sayeed, M. (2014). Design and implementation of butterworth, chebyshev-i and elliptic filter for speech signal analysis, *International Journal of Computer Applications* **98**(7).

Pratt, R. and Doak, P. (1976). A subjective rating scale for timbre, *Journal of Sound and Vibration* **45**(3): 317–328.

Pressnitzer, D. and McAdams, S. (2000). Acoustics, psychoacoustics and spectral music, *Contemporary Music Review* **19**(2): 33–59.

Ramo, J., Valimaki, V. and Bank, B. (2014). High-precision parallel graphic equalizer, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **22**(12): 1894–1904.

Read, O. (1952). *The recording and reproduction of sound*, Indianapolis: HW Sams.

Reed, D. (2000). A perceptual assistant to do sound equalization, *Proceedings of the 5th international conference on Intelligent user interfaces*, ACM, pp. 212–218.

Rossing, T. D., Moore, F. R., Wheeler, P. A. and Rossing-Moore-Wheeler (2002). *The science of sound*, Vol. 3, Addison Wesley San Francisco.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of computational and applied mathematics* **20**: 53–65.

Roweis, S. T. (1998). Em algorithms for pca and spca, *Advances in neural information processing systems*, pp. 626–632.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding, *science* **290**(5500): 2323–2326.

Roweis, S. et al. (2002). Automatic alignment of hidden representations, *Sixteenth Annual Conference on Neural Information Processing Systems, Vancouver, Canada*, Vol. 15, pp. 841–848.

Rubio, J. P. A. (2005). *An object-oriented metamodel for digital signal processing with a focus on audio and music*, PhD thesis, Universitat Pompeu Fabra.

Rumelhart, D. E., Hinton, G. E., Williams, R. J. et al. (1988). Learning representations by back-propagating errors, *Cognitive modeling* **5**(3): 1.

Rumsey, F. (2005). Psychoacoustics of sound quality, *Proceedings of the Art of Record Production Conference, London*.

Sabin, A. and Pardo, B. (2008). Rapid learning of subjective preference in equalization, *Audio Engineering Society Convention 125*, Audio Engineering Society.

Sabin, A. T. and Pardo, B. (2009a). 2DEQ: An intuitive audio equalizer, *Proceedings of the seventh ACM conference on Creativity and cognition*, ACM, pp. 435–436.

Sabin, A. T. and Pardo, B. (2009b). A method for rapid personalization of audio equalization parameters, *Proceedings of the 17th ACM international conference on Multimedia*, ACM, pp. 769–772.

Sabin, A. T., Rafii, Z. and Pardo, B. (2011). Weighted-function-based rapid mapping of descriptors to audio processing parameters, *Journal of the Audio Engineering Society* **59**(6): 419–430.

Saitis, C., Giordano, B. L., Fritz, C. and Scavone, G. P. (2012). Perceptual evaluation of violins: A quantitative analysis of preference judgments by experienced players, *The Journal of the Acoustical Society of America* **132**(6): 4002–4012.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis, *IEEE Transactions on computers* **100**(5): 401–409.

Sanguinetti, G. (2008). Dimensionality reduction of clustered data sets, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**(3): 535–540.

Sankiewicz, M. and Budzyński, G. (2007). Reflections on sound timbre definitions, *Archives of Acoustics* **32**(3): 591–602.

Saraçli, S., Doğan, N. and Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation, *Journal of Inequalities and Applications* **2013**(1): 203.

Schölkopf, B., Smola, A. and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural computation* **10**(5): 1299–1319.

Seetharaman, P. and Pardo, B. (2014a). Crowdsourcing a reverberation descriptor map, *Proceedings of the 22nd ACM international conference on Multimedia*, ACM, pp. 587–596.

Seetharaman, P. and Pardo, B. (2014b). Reverbalize: a crowdsourced reverberation controller, *Proceedings of the 22nd ACM international conference on Multimedia*, ACM, pp. 739–740.

Sheikholeslami, G., Chatterjee, S. and Zhang, A. (1998). Wavecluster: A multi-resolution clustering approach for very large spatial databases, *VLDB*, Vol. 98, pp. 428–439.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. i., *Psychometrika* **27**(2): 125–140.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation, *IEEE Transactions on pattern analysis and machine intelligence* **22**(8): 888–905.

Sibson, R. (1981). A brief description of natural neighbour interpolation, *Interpreting multivariate data* .

Slaney, M., Covell, M. and Lassiter, B. (1996). Automatic audio morphing, *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, Vol. 2, IEEE, pp. 1001–1004.

Slawson, W. (1985). *Sound color*, Yank Gulch Music.

Smith, S. W. et al. (1997). *The scientist and engineer's guide to digital signal processing*, California Technical Pub. San Diego.

Solomon, L. N. (1959). Search for physical correlates to psychological dimensions of sounds, *The Journal of the Acoustical Society of America* **31**(4): 492–497.

Stables, R., De Man, B., Enderby, S., Reiss, J. D., Fazekas, G. and Wilmering, T. (2016). Semantic description of timbral transformations in music production, *Proceedings of the 2016 ACM on Multimedia Conference*, ACM, pp. 337–341.

Stables, R., Enderby, S., Man, B., Fazekas, G., Reiss, J. D. et al. (2014). Safe: A system for the extraction and retrieval of semantic audio descriptors.

Stasis, S., Hockman, J. and Stables, R. (2016). Descriptor sub-representations in semantic equalisation, *Proceedings of the 2nd AES Workshop on Intelligent Music Production*, Vol. 13.

Stasis, S., Hockman, J. and Stables, R. (2017). Navigating descriptive sub-representations of musical timbre, *Proceedings of the International Conference on New Interfaces for Musical Expression, Copenhagen, Denmark*.

Stasis, S., Jillings, N., Enderby, S. and Stables, R. (2017a). Audio processing chain recommendation, *Proceedings of the 20th International Conference on Digital Audio Effects,(Edinburgh, UK)*.

Stasis, S., Jillings, N., Enderby, S. and Stables, R. (2017b). Audio processing chain recommendaton using semantic cues, *Proceedings of the 3rd Workshop on Intelligent Music Production*.

Stasis, S., Stables, R. and Hockman, J. (2015). A model for adaptive reduced-dimensionality equalisation, *18th International International Conference on Digital Audio Effects (DAFx-15), Trondheim, Norway*.

Stasis, S., Stables, R. and Hockman, J. (2016). Semantically controlled adaptive equalisation in reduced dimensionality parameter space, *Applied Sciences* **6**(4): 116.

Stepánek, J. (2006). Musical sound timbre: Verbal description and dimensions, *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pp. 121–126.

Tan, P.-N. et al. (2006). *Introduction to data mining*, Pearson Education India.

Tassart, S. (2013). Graphical equalization using interpolated filter banks, *Journal of the Audio Engineering Society* **61**(5): 263–279.

Tauchen, G. (1986). Finite state markov-chain approximations to univariate and vector autoregressions, *Economics letters* **20**(2): 177–181.

Tenenbaum, J. B., De Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction, *science* **290**(5500): 2319–2323.

Terhardt, E. (1974). On the perception of periodic sound fluctuations (roughness), *Acta Acustica united with Acustica* **30**(4): 201–213.

Theodoridis, S., Pikrakis, A., Koutroumbas, K. and Cavouras, D. (2010). *Introduction to pattern recognition: a matlab approach*, Academic Press.

Thomas, L. (1971). The biquad: Part I - some practical design considerations, *IEEE Transactions on Circuit Theory* **18**(3): 350–357.

Thomson, W. (1949). Delay networks having maximally flat frequency characteristics, *Proceedings of the IEE-Part III: Radio and Communication Engineering* **96**(44): 487–490.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude, *COURSERA: Neural networks for machine learning* **4**(2): 26–31.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method, *Psychometrika* **17**(4): 401–419.

Toulson, E. (2006). A need for universal definitions of audio terminologies and improved knowledge transfer to the audio consumer, *Presented at The Art of Record Production Conference*, Vol. 8, p. 10.

Tow, J. (1968). Active rc filters—a state-space realization, *Proceedings of the IEEE* **56**(6): 1137–1139.

Traube, C., Bernays, M. and Bellemare, M. (2008). Perception, verbal description and gestural control of piano timbre, *The Journal of the Acoustical Society of America* **123**(5): 3657–3657.

Van der Maaten, L. (2007). An introduction to dimensionality reduction using matlab, *Report* **1201**(07-07): 62.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne, *Journal of Machine Learning Research* **9**(Nov): 2579–2605.

Van Der Maaten, L., Postma, E. and Van den Herik, J. (2009). Dimensionality reduction: a comparative review, *J Mach Learn Res* **10**: 66–71.

Van Waterschoot, T. and Moonen, M. (2011). Fifty years of acoustic feedback control: State of the art and future challenges, *Proceedings of the IEEE* **99**(2): 288–327.

Venna, J. and Kaski, S. (2005). Local multidimensional scaling with controlled tradeoff between trustworthiness and continuity, *Proceedings of WSOM*, Vol. 5, Citeseer, pp. 695–702.

Venna, J. and Kaski, S. (2006). Visualizing gene interaction graphs with local multidimensional scaling., *ESANN*, Vol. 6, pp. 557–562.

Verfaille, V. and Arfib, D. (2000). A-dafx: Adaptive digital audio effects, *energy* **4000**: 6000.

Verfaille, V., Zolzer, U. and Arfib, D. (2006). Adaptive digital audio effects (a-dafx): A new class of sound transformations, *IEEE Transactions on audio, speech, and language processing* **14**(5): 1817–1831.

Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders, *Proceedings of the 25th international conference on Machine learning*, ACM, pp. 1096–1103.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, *Journal of Machine Learning Research* **11**(Dec): 3371–3408.

von Bismarck, G. (1974). Timbre of steady sounds: A factorial investigation of its verbal attributes, *Acta Acustica united with Acustica* **30**(3): 146–159.

von Helmholtz, H. (1912). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, Longmans, Green.

von Luxburg, U. (2007). A tutorial on spectral clustering, *Statistics and computing* **17**(4): 395–416.

Vurma, A., Raju, M. and Kuuda, A. (2011). Does timbre affect pitch?: Estimations by musicians and non-musicians, *Psychology of Music* **39**(3): 291–306.

Waddell, G. (2013). *Complete Audio Mastering: Practical Techniques*, McGraw-Hill Education.

Wake, S. and Asahi, T. (1998). Sound retrieval with intuitive verbal expressions, Georgia Institute of Technology.

Wessel, D. L. (1973). Psychoacoustics and music: A report from michigan state university, *PACE: Bulletin of the Computer Arts Society* **30**: 1–2.

Wessel, D. L. (1979). Timbre space as a musical control structure, *Computer music journal* pp. 45–52.

White, P. (1999). *Basic Mixers*, The Basic Series, Music Sales.

Wickelmaier, F. (2003). An introduction to mds, *Sound Quality Research Unit, Aalborg University, Denmark* **46**.

Williams, D. and Brookes, T. (2009). Perceptually-motivated audio morphing: softness, *Audio Engineering Society Convention 126*, Audio Engineering Society.

Winsberg, S. and Carroll, J. D. (1989). A quasi-nonmetric method for multidimensional scaling via an extended euclidean model, *Psychometrika* **54**(2): 217–229.

Winsberg, S. and De Soete, G. (1993). A latent class approach to fitting the weighted euclidean model, clascal, *Psychometrika* **58**(2): 315–330.

Wise, D. K. (2009). Concept, design, and implementation of a general dynamic parametric equalizer, *Journal of the Audio Engineering Society* **57**(1/2): 16–28.

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

Zacharakis, A., Pastiadis, K., Papadelis, G. and Reiss, J. D. (2011). An investigation of musical timbre: Uncovering salient semantic descriptors and perceptual dimensions., *ISMIR*, pp. 807–812.

Zacharakis, A. and Reiss, J. (2011). An additive synthesis technique for independent modification of the auditory perceptions of brightness and warmth, *Audio Engineering Society Convention 130*, Audio Engineering Society.

Zhang, Z. and Zha, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment, *SIAM journal on scientific computing* **26**(1): 313–338.

Zheng, T., Seetharaman, P. and Pardo, B. (2016). Socialfx: Studying a crowdsourced folksonomy of audio effects terms, *Proceedings of the 2016 ACM on Multimedia Conference*, ACM, pp. 182–186.

Zobel, O. J. (1923). Theory and design of uniform and composite electric wave-filters, *Bell Labs Technical Journal* **2**(1): 1–46.

Zölzer, U. and Smith, J. O. (2003). Dafx—digital audio effects, *The Journal of the Acoustical Society of America* **114**(5): 2527–2528.

Zumbahlen, H. et al. (2011). *Linear circuit design handbook*, Newnes.