Musical timbre: bridging perception with semantics.

Asterios Zacharakis

Submitted to the University of London in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Queen Mary University of London

2013

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline. I acknowledge the helpful guidance and support of my supervisor, Joshua D. Reiss.

Abstract

Musical timbre is a complex and multidimensional entity which provides information regarding the properties of a sound source (size, material, etc.). When it comes to music, however, timbre does not merely carry environmental information, but it also conveys aesthetic meaning. In this sense, semantic description of musical tones is used to express perceptual concepts related to artistic intention. Recent advances in sound processing and synthesis technology have enabled the production of unique timbral qualities which cannot be easily associated with a familiar musical instrument. Therefore, verbal description of these qualities facilitates communication between musicians, composers, producers, audio engineers etc. The development of a common semantic framework for musical timbre description could be exploited by intuitive sound synthesis and processing systems and could even influence the way in which music is being consumed.

This work investigates the relationship between musical timbre perception and its semantics. A set of listening experiments in which participants from two different language groups (Greek and English) rated isolated musical tones on semantic scales has tested semantic universality of musical timbre. The results suggested that the salient semantic dimensions of timbre, namely: *luminance*, *texture* and *mass*, are indeed largely common between these two languages. The relationship between semantics and perception was further examined by comparing the previously identified semantic space with a perceptual timbre space (resulting from pairwise dissimilarity rating of the same stimuli). The two spaces featured a substantial amount of common variance suggesting that semantic description can largely capture timbre perception. Additionally, the acoustic correlates of the semantic and perceptual dimensions were investigated. This work concludes by introducing the concept of partial timbre through a listening experiment that demonstrates the influence of background white noise on the perception of musical tones. The results show that timbre is a relative percept which is influenced by the auditory environment.

Acknowledgements

I would like to thank all the people that I met and worked with during my Ph.D study in the Centre for Digital Music at Queen Mary University of London. Especially my supervisor, Josh Reiss, for trusting me and giving me the freedom to follow all the research paths that I believed were worth pursuing.

I would also like to thank Enrique Perez-Gonzalez, Andrew Robertson, Jean-Baptiste Thiebaut and Dan Stowell for the fruitful conversations and guidance especially in the early uncertain times of my study. Thanks to Manolis Benetos for his most significant help with LaTeX. Also, thanks to Marcus Pearce, Mathieu Barthet, Georgios Papadelis, Marcelo Caetano, Petri Toiviainen and Andrea Halpern for providing valuable feedback on various parts of my work. Special thanks to Robin Hart, Sussan Sturrock and John Dack for helping me organise a number of listening experiments at the Royal College of Music and Middlesex University. I would also like to acknowledge the contribution of Stephen McAdams, Bruno Giordano and Jeremy Marozeau to shaping my work through their detailed peer reviewing.

At this point, I have to single out three very special colleagues without whom this work would not have been the same; Andy Simpson and Mike Terrell from C4DM and Kostas Pastiadis, lecturer at the Department of Music Studies at the Aristotle University of Thessaloniki. It is hard to describe their contribution to my work in just a few sentences. Andy and Mike's vivid interest in my work has allowed them to gain remarkable insight into it and has acted as a strong motivation for me. Andy has taught me how to better present my work in writing and his resilience in all sort of tough situations has been a great source of inspiration and strength. Mike's experience in doing research (already holding two Ph.D titles!) and high expertise, both of which he profusely offered at all times, have been proven extremely valuable at many instances. I have one extra reason to be grateful to Andy and Mike, since their families have made London feel more like home to me. Finally, Kostas' enthusiasm for the core idea of my work along with his broad knowledge on the field of music acoustics and perception have constituted him an ideal collaborator. Kostas has actually acted as a co-supervisor of my Ph.D by widely offering his time, energy and

resources despite the deteriorating environment for academic research in crisis-stricken Greece. I am grateful to you all and privileged to be your friend.

A big thanks to all my dear friends for their support. I am especially grateful to uncle Yannis, Veneta and Eleni for taking good care of me throughout my four London years. Thanks to my friend Yannis for the maths, the programming, the discussions, the delicious pasta and for letting me win at chess from time to time. To mum, dad and Fivos, thank you, for your love and support throughout my life. None of these could have been possible without you. Finally, I would like to thank my partner Lenia, for enduring a three-year distant relationship for the sake of my Ph.D study. Thank you for fighting your fear of planes, for your sleepless nights at the airports, for being so caring and understanding and for encouraging me through excitements or disappointments. I love you.

Contents

List of Figures

List of Tables

"*The trombones crunched redgold under my bed, and behind my gulliver the trumpets three-wise silverflamed, and there by the door the timps rolling through my guts and out again crunched like candy thunder. Oh, it was wonder of wonders. And then, a bird of like rarest spun heaven-metal, or like silvery wine flowing in a spaceship, gravity all nonsense now, came the violin solo above all the other strings, and those strings were like a cage of silk around my bed. Then flute and oboe bored, like worms of like platinum, into the thick thick toffee gold and silver. I was in such bliss, my brothers.*"

Anthony Burgess, *Clockwork Orange* [Burgess, 1986]

Chapter 1

Introduction

1.1 Motivation

Musical timbre is perhaps the most complex and fascinating attribute of sound. It plays a very important role for sound identification but also for defining the aesthetic quality of a sound object, that in turn is crucial for music appreciation. Semantic description of timbre through verbal means is quite common, especially among musicians. Interestingly enough, the rich vocabulary used for timbre description sometimes resembles the vocabulary used for description of other aesthetic objects such as alcoholic drinks. Lexical description of wines for example [Lehrer, 2007], can be useful for advertising the qualities of a product and facilitating selection or simply for the pleasure of communication while wine tasting. But how is timbre description useful? Of course, communicating through verbal means regarding a listening experience can also be pleasurable but can timbre semantics really influence music creation or appreciation?

Before the development of computers, the options for timbre manipulation were limited by the available instruments and their combinations. Composers could, of course, push back the existing timbral frontiers by requesting novel playing techniques or by utilising the art of orchestration so as to produce interesting sonic combinations. However, it was not until the development of electric and electronic instruments, only a few decades ago, that the available timbral palette was vastly enriched. It would not be an overstatement to suggest that these technological advances have essentially enabled the creation of any imaginable timbre. It was not always easy for musicians to follow the technological innovations and as a result, technologically qualified

individuals such as audio engineers became a significant factor in popular music creation.

Despite the fact that many musicians are increasingly developing the necessary technical skills, they still often delegate part of their vision to producers or audio engineers who, among other things, act as a bridge between available technological means and artistic intentions. This apparently requires a description of intention by the artist. John Lennon, for example, was particularly fond of intuitively describing how he envisioned his songs. In an indicative anecdote it is said that he had once asked producer George Martin to make one of his tracks 'sound like an orange'. Furthermore, his request of a 'fairground' production wherein someone could smell the sawdust [MacDonald, 1995, p. 210] for 'Being for the benefit of Mr Kite!' resulted in the known brilliant arrangement. Obviously, the producer had to map an abstract, high level description of artistic intention into something musically relevant and timbre manipulation (of single instruments or of the whole mix) certainly offers one possible way to satisfy such a description. Timbral descriptions can be particularly useful in an era where novel timbres are highly available. Potential applications of timbral semantics include sound synthesis, music production and reproduction, music education, sound design, etc.

1.2 Thesis objectives

The main objective of this work is to establish a common semantic framework for describing the timbre of musical tones. To this end, this thesis will investigate three fundamental questions:

1. Are timbre semantics universal or do they depend on language? The influence of language on timbre semantics will be examined through the comparison of semantic spaces resulting from English and Greek verbal descriptions.

2. What is the relationship of semantics with perception? The semantic and perceptual spaces for the same set of sounds will be compared to test the amount of perceptual information that can be conveyed through semantic description.

3. Finally, is the timbre of a sound an absolute percept or is it influenced by the auditory environment? An initial exploration of the influence of the acoustic environment on timbre perception concludes this work.

The answers to the above questions will define whether the development of a common semantic framework for timbre is feasible and meaningful.

1.3 Thesis Structure

Chapter 2 presents the background of timbre perception studies. It starts by introducing the concept of timbre together with the main experimental approaches for its investigation and proceeds with the main acoustic correlates. It subsequently presents the background on timbre semantics and discusses their relationship with perception. The chapter concludes by discussing interdependencies of timbre with pitch or with the auditory environment.

Chapter 3 provides a description of the basic statistic tools employed for the purposes of this work along with a presentation of the full set of acoustic descriptors that were extracted and investigated.

Chapter 4 describes an initial attempt to investigate timbral semantics (auditory *brightness* and *warmth* in particular) and their acoustic correlates. The unclear results of this very specific experiment have demonstrated the need to adopt a more holistic approach. The conclusion of this chapter discusses the identified weaknesses and how they were addressed by the subsequent experimental design.

Chapter 5 presents a listening experiment that tested the universality of musical timbre semantics and identified the acoustic correlates of the salient semantic dimensions. Native Greek and English participants took part in two separate timbre description experiments and the results of each language group were discussed and compared. The analysis has additionally accounted for potential nonlinear relationships between the semantic variables which resulted in more robust semantic spaces.

Chapter 6 extends the findings of chapter 5. The previously identified semantic space was compared with a perceptual space that resulted from a pairwise dissimilarity listening test and did not involve any semantic description. The similarities between the two spaces indicated that semantic description of timbre is capable of conveying perceptual information. The acoustic correlates of the perceptual dimensions were also found to be largely similar with the semantic space ones.

Chapter 7 introduces the concept of *partial timbre* for describing the portion of the original timbre (i.e. timbre in isolation) that remains in a sound when heard in a complex auditory scene. A series of pairwise dissimilarity listening tests were conducted on the same set of

harmonic sounds under three different background noise conditions. The findings revealed that the perceptual structure of a set of sounds is significantly affected by the level of background noise.

Chapter 8 summarises the major contributions of this thesis and proposes fruitful areas for future work.

All the listening experiments that were conducted for the purposes of this thesis were approved by the Research Ethics Committee at Queen Mary University of London. Listening tests were run on an ad-hoc basis and participants gave verbal, informed consent. Participants were also free to withdraw at any point.

1.4 Associated Publications

Portions of the work presented in this thesis have been published in various international scholarly publications, as follows:

- The largest part of chapter 4 was presented at the 130th Audio Engineering Society Convention [Zacharakis and Reiss, 2011].
- Chapter 5 is a more detailed version of a paper accepted for publication in *Music Perception* [Zacharakis et al., accepted]. Additionally, portions of this chapter have been presented at the International Society for Music Information Retrieval (ISMIR) conference [Zacharakis et al., 2011] and at the joint conference of the International Conference on Music Perception and Cognition and the Trienial Conference of the European Society for the Cognitive Sciences of Music (ICMPC-ESCOM) [Zacharakis et al., 2012].
- Chapter 6 has been submitted for publication to *Music Perception*.
- Chapter 7 presents a collaborative study of which I was the lead author and which has been submitted for publication to *PLOS ONE*.

Chapter 2

Musical Timbre

2.1 A problematic definition

The investigation of musical timbre perception has a long history. von Helmholtz [1877] set the foundations of acoustics and sound perception at the end of the 19th century. However, it was not until the early seventies that timbre perception research began to flourish. Timbre is regarded as one of the four major auditory attributes of tone, the rest being loudness, pitch and duration[1]. Out of the four, timbre is by far the most complex attribute, featuring both categorical and continuous characteristics. Additionally, its multidimensional nature is evidently influenced by loudness, pitch and duration, making it hard to even come up with a solid definition.

The ANSI [1973] definition, according to which *timbre is that attribute of auditory sensation in terms of which a subject can judge that two sounds similarly presented and having the same loudness, pitch and duration are dissimilar*, is a definition by negation. As such, it has been criticised by various researchers [e.g. Sankiewicz and Budzynski, 2007, Donnadieu, 2007, Papanikolaou and Pastiadis, 2009] but nevertheless a really alternative definition has not yet been suggested. Albert Bregman [1994], one of the most prominent researchers in the field of auditory perception and 'father' of Auditory Scene Analysis (ASA), has stated that the ANSI definition

> "... is, of course, no definition at all. For example, it implies that there are some sounds for which we cannot decide whether they possess the quality of timbre or not. In order for the definition to apply, two sounds need to be able to be presented

[1] Some researchers additionally include spatial position.

> at the same pitch, but there are some sounds, such as the scarping of a shovel in a pile of gravel, that have no pitch at all. We obviously have a problem: Either we must assert that only sounds with pitch can have timbre, meaning that we cannot discuss the timbre of a tambourine or of the musical sounds of many African cultures, or there is something terribly wrong with the definition."

The refined definition by Pratt and Doak [1976]:

> "Timbre is that attribute of auditory sensation whereby a listener can judge that two sounds are dissimilar using any criteria other than pitch, loudness or duration."

has bypassed the pitch issue highlighted by Bregman but is, nevertheless, rarely cited as a timbre definition. According to this definition a sound does not necessarily need to have a clear pitch in order to possess timbre. Krumhansl [1989] attributes the difficulty to reach a general definition of timbre to the fact that it is so closely associated with the set of traditional orchestral instruments.

The basic sources of criticism regarding the ANSI definition of timbre sum up to the following points: (1) not all musical sounds feature a clear-cut pitch or static loudness, (2) timbre may refer to concurrent sounding tones of different instruments or to a complex sound structure, (3) timbre may also refer to a specific element of the sound object such as the attack.

2.2 Classification and relational measures

In an effort to isolate timbre, researchers initially considered single isolated synthesised or acoustic tones that were equalised for loudness, pitch and perceived duration (for an overview on timbre perception studies see Hajda [2007] and Donnadieu [2007]). According to Hajda [2007], all methods in the timbre perception literature that target to formulate groups of objects such as categorisation, recognition or identification fall under the broad term *classification*. The recognition and identification of sound sources is arguably the most significant task of our auditory system in evolutionary terms and when it comes to musical timbre we are able to identify specific musical instruments and instrument families. Smalley [1997] describes our natural tendency to relate sounds to supposed sources or causes (actual or imagined) with the term *source bonding*. We are, additionally, capable of associating a range of varying timbres with a single instrument (e.g. *sul tasto* vs *sul ponticello* or *con legno* playing techniques in bowed strings).

The second set of methods that are often utilised by timbre perception studies aim at comparison between sound objects through interval or ratio measures. Hajda [2007] calls this approach *relational measures*. The basic representative of direct relational measures is pairwise dissimilarity rating where pairs of sounds are directly compared for similarity [e.g. Plomp, 1970, 1976, Miller and Carterette, 1975, Iverson and Krumhansl, 1993, Caclin et al., 2005]. An indirect way to measure the relationships between sounds is through *verbal attribute magnitude estimation* (VAME) [Kendall and Carterette, 1993a,b] or *semantic differential* [von Bismarck, 1974a]. These methods require the rating of sound objects along semantic scales and will be further discussed below.

2.3 Multidimensional Scaling analysis and timbre spaces

In the early seventies a new statistical tool was introduced to the study of timbre perception. This tool was Multidimensional Scaling (MDS) analysis and was initially utilised in timbre research by Plomp [1970]. MDS originates from psychometrics and was developed to enable the interpretation of people's pairwise dissimilarity judgements over a set of perceptual objects [Shepard, 1962a,b]. The various MDS algorithms produce N-dimensional geometric configurations (and inform about their optimal dimensionality) based on maximising the goodness-of-fit measures that relate Euclidean distances between points in the space to the actual dissimilarity ratings between perceptual objects.

Following the influential work by Grey [1977], the MDS approach has become a norm for timbre perception investigation [e.g. Kendall and Carterette, 1991, Iverson and Krumhansl, 1993, McAdams et al., 1995, Caclin et al., 2005] because of its ability to construct low dimensional spatial representations of the perceptual objects under study, a desirable property for the investigation of complex entities. In the case of timbre, these constructs are called *timbre spaces* and offer visualisation of the perceptual structure within a set of sounds. Thus, they are particularly useful for the identification of the salient perceptual dimensions of timbre (i.e., dimensions that best explain the perceived dissimilarities between the stimuli). Previous studies on the perception of musical timbre have identified either 3 or 4 major perceptual dimensions for modelling timbres of monophonic acoustic instruments [e.g. Grey, 1977, Krimphoff, 1993, Krimphoff et al., 1994, McAdams et al., 1995]. Figure 2.1 shows an example of a timbre space.

McAdams [1999] offers an overview of available MDS techniques along with their use in

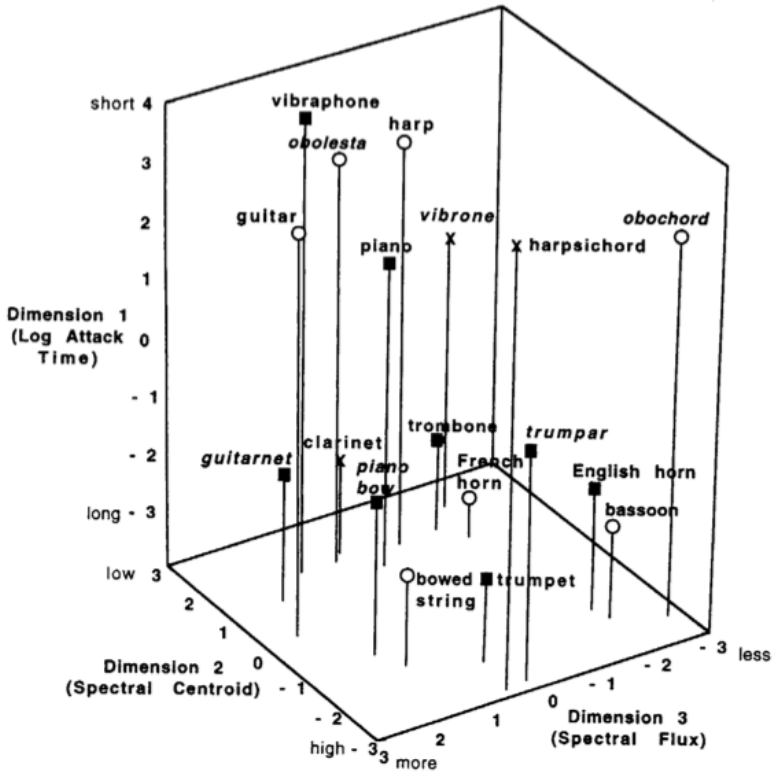


Figure 2.1: An example of a three-dimensional timbre space from McAdams et al. [1995].

timbre perception research. MDS in its classical form was designed to interpret a single set of dissimilarities among items and not the average over all participants of an experiment. Initially, distance models were either Euclidean or Minkowski generalizations of the Euclidean distance. According to these models the distance d_{ij} between any two timbres i and j is given by Equation 2.1:

$$d_{ij} = [\sum_{k=1}^{K} (x_{ik} - x_{jk})^r]^{1/r} \tag{2.1}$$

where x_{ik} is the coordinate of timbre i on dimension k, K represents the total number of dimensions in the model and r determines the Minkowski metric. The norm for timbre studies is a Euclidean distance model, i.e. $r = 2$, which produces a Euclidean timbre space under the condition that the number of examined timbres is much larger than the number of dimensions[2]. This model was utilised by some early studies [Wessel, 1973, 1979] through the MDSCAL program

[2]A commonly used rule of thumb is that at least four stimuli are required per dimension [Green et al., 1989]. This means that the minimum number of stimuli for obtaining a 3D perceptual space should be twelve.

Kruskal [1964a,b]. However, this model presumes that the set of dimensions is not only common for each listener but that the dimensions are also equally weighted perceptually. This seems like an unjustified hypothesis since we know that listeners are not equally sensitive to every auditory parameter [McAdams et al., 1995].

In order to address this effect, the spatial model has been extended to the following form:

$$d_{ij} = [\sum_{k=1}^{K} w_{nk}(x_{ik} - x_{jk})^2]^{1/2} \tag{2.2}$$

where w_{nk} represents the perceptual 'weight' [0-1] attributed to dimension k by listener n. The above model was realised by the INDSCAl program [Carroll and Chang, 1970] and has been used by a number of studies [e.g. Miller and Carterette, 1975, Grey, 1977, Grey and Gordon, 1978]. However, the separate treatment of each listener drastically increases the parameters of the model as a consequence of increasing the number of participants. To alleviate this issue, it has been proposed that listeners are not treated individually, but as part of a small number of 'latent classes' that represent groups of listeners who pursue similar rating strategies. Thus, the individual weights are replaced by weights for each class of participants. Based on statistical tests on the data, the probability that each listener belongs to each class is calculated and class membership is assigned to each participant accordingly. This approach was implemented by the CLASCAL algorithm [Winsberg and Soete, 1993].

Both of the above models are based on the hypothesis that all of the variance in a data set can be explained by dimensions common to all stimuli. However, it seems probable that some of the sounds may feature unique characteristics, not shared by the rest of the stimuli in the set, that can be perceptually significant. Such 'specificities' would certainly contribute to dissimilarities between sounds but cannot be accounted for by the common continuous dimensions of a timbre space. Therefore, another type of distance model extension was suggested based on the following Equation:

$$d_{ij} = [\sum_{k=1}^{K} (x_{ik} - x_{jk})^2 + s_i + s_j]^{1/2} \tag{2.3}$$

where s_i and s_j are the specificities corresponding to timbres i and j respectively. A specificity can either represent the coordinate $\sqrt{s_i}$ along an additional dimension on which only timbre i varies or it can represent the perceptual salience of a discrete feature present only in timbre x_i.

This extended model was implemented by the EXSCAL program [Winsberg and Carroll, 1989].

Finally, a combination of the 'latent classes' and specificities approaches has led to a model that incorporates both class weights and specificities. As shown in Equation 2.4, weights are not only applied to both continuous dimensions but also to the whole set of specificities:

$$d_{ij} = [\sum_{k=1}^{K} (w_{kc}(x_{ik} - x_{jk})^2 + v_c(s_i + s_j)]^{1/2} \tag{2.4}$$

where w_{kc} is the weight on dimension k for class c and v_c is the weight on the set of specificities. This combination of the CLASCAL and EXSCAL models was used in one of the most comprehensive studies on timbre perception by McAdams et al. [1995].

As stated in Burgoyne and McAdams [2007], one potential issue with all the previously presented MDS techniques is that, being linear, they consider all distances estimated by the human subjects as being equally reliable and of equally relative scale. In a meta-analysis of data from Grey [1977], Grey and Gordon [1978], McAdams et al. [1995] with a nonlinear extension of MDS, Burgoyne and McAdams [2007, 2008] showed that a nonlinear treatment of pairwise dissimilarity ratings can preserve the spatial structure with fewer dimensions[3]. This implied that the nonlinearities present in timbre judgements are significant and should be considered in the analysis.

Figure 2.2 presents the steps that usually constitute a pairwise dissimilarity rating experiment. MDS analysis is followed by the physical interpretation of the identified dimensions. In the case of *automated timbre spaces* [Nicol, 2005] where each sound is represented by a vector of acoustic features [e.g. Hourdin et al., 1997] this interpretation is direct. In the case of *human timbre spaces* [Nicol, 2005], however, where the dissimilarities between sounds come from human judgements, the physical interpretation of the dimensions is achieved by computing the correlations between the positions of the sounds on each dimension with the extracted acoustic descriptors. Collections of acoustic descriptors that are widely used in timbre perception literature are presented in Peeters [2004] and Peeters et al. [2011]. More specifically Peeters et al. [2011] discuss a comprehensive set of audio descriptors that are calculated by the matlab Timbre Toolbox. The next section will present the most prominent acoustic correlates of timbral perceptual dimensions.

[3]The strategy adopted was the preprocessing of dissimilarity matrices with the nonlinear Isomap algorithm [Tennenbaum et al., 2000] which were subsequently fed into the CLASCAL algorithm. The Isomap transformation emphasises the effect of smaller differences between timbres that are perceived as fairly similar and reduces the effect of large differences between distant timbres.

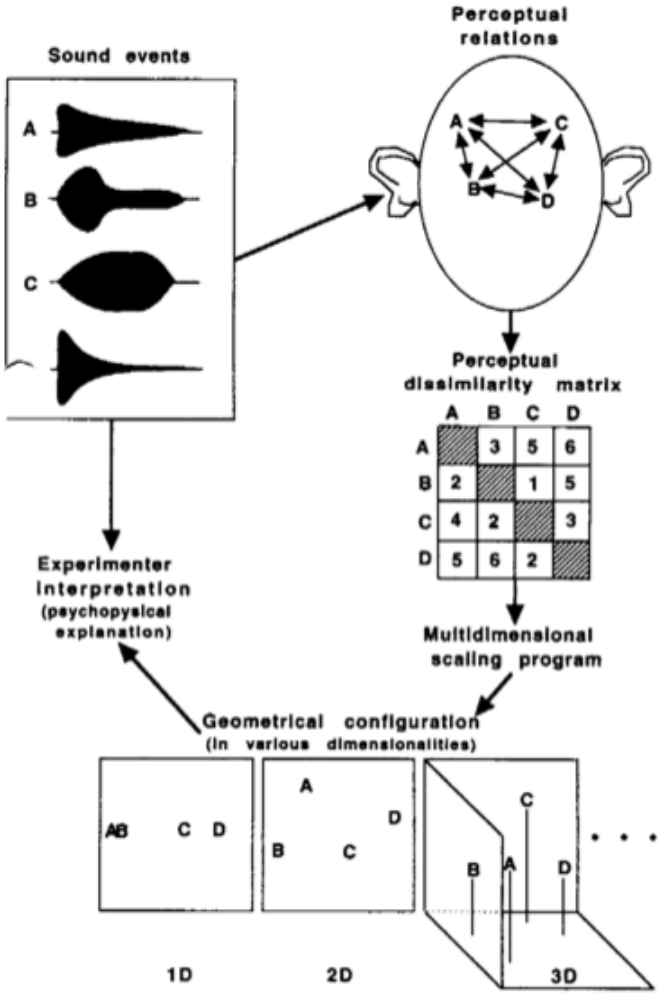


Figure 2.2: A pairwise dissimilarity experiment consists of the following steps: pairwise dissimilarity rating → perceptual dissimilarity matrix → MDS analysis → timbre space → psychophysical interpretation. From McAdams [1999].

2.4 Acoustic correlates

The first one to propose that timbre of a sound was dependent on the amplitudes of frequencies present in the sound was Ohm [1843]. von Helmholtz [1877] has also proposed a set of rules for associating semantic descriptions of musical timbre with acoustic properties a summary of which can be found in Howard and Tyrrell [1997]. The development of computers during the last fifty years has enabled the calculation of acoustic descriptors from sound signals and has (in combination with the MDS approach) facilitated the association of identified perceptual dimensions with physical properties of musical sound. Knopoff [1963, p.29] was the first to introduce the barycentre of the spectrum (spectral centroid) as a measure of musical instrument quality. Spectral centroid was first correlated with a perceptual dimension of timbre by Ehresman and

Wessel [1978] and Grey and Gordon [1978]. Grey [1977] in his classic paper has proposed acoustic correlates for his three identified perceptual dimensions of musical timbre. Since then, a plethora of studies have investigated acoustic correlates of synthesised or natural tones arriving at conclusions that are not always consistent.

The comprehensive review of Peeters et al. [2011] has organised audio descriptors according to three of their main properties:

1. The *time extent* over which the descriptor is computed. This can either be the whole signal or a segment duration. In the first case the descriptor is called *global* and characterises the whole sound event. Examples of *global descriptors* are the logarithm of the attack time (as there is only one attack in a tone), temporal centroid, effective duration etc. When the descriptor is calculated from a time frame (i.e., very short segment) of the signal it is called *time-varying* or *instant*. *Time varying descriptors* consist of a sequence of values each of which corresponds to a separate time frame. Thus, they are usually further treated by means of descriptive statistics (e.g. maximum and minimum values, mean, median, standard deviation, interquartile range etc.) so as to obtain a single value which will represent the whole sound sample.

2. The *signal representation* used to compute the descriptor. There are various available input signal representations that are often utilised. The waveform or the energy envelope are mostly used for computing temporal or spectrotemporal descriptors. Transformations of the signal such as the short term Fourier transform (magnitude and power scale) or the wavelet transform are often used for calculating spectral descriptors while a sinusoidal modeling output is used for calculating harmonic features (e.g. inharmonicity, harmonic spectral centroid, etc.). Representations that try to mimic the output of the middle ear (i.e., based on an auditory model) such as the bark-scale or the Equivalent Rectangular Bandwidth (ERB) filter banks can alternatively be used.

3. The *concept of the descriptor* refers to the particular aspect of the sound signal that is being measured by the descriptor regardless of its input representation. A descriptor can represent the spectral envelope (e.g. spectral centroid, spectral spread, spectral slope, spectral roll-off, etc.), the temporal envelope (e.g. logarithm of the attack time, temporal centroid, etc.), harmonic characteristics (e.g. harmonic spectral centroid, inharmonicity, odd to even harmonic ratio, tristimulus, etc.), energy content (e.g. global energy, harmonic energy, noise energy,

etc.), spectrotemporal characteristics (e.g. spectral variation, mean coefficiant of variation, etc.)

While it is generally accepted that spectral envelope, temporal envelope and spectrotemporal variations influence timbre perception significantly, there does not exist an absolute consensus regarding acoustic correlates. It seems that both the selection of stimuli and the variations in the definitions and calculations of acoustic descriptors contribute towards this ambiguity. It can also be argued that trying to represent a dynamic entity such as timbre by a single numerical value can also pose a problem. Furthermore, a study on synthetic tones by Caclin et al. [2005] provided evidence that the salience of some acoustic descriptors is context dependent. They particularly showed that the perceptual significance of a spectrotemporal characteristic like spectral flux decreases when the number of acoustic parameters (e.g. spectral centroid and attack time) that concurrently vary in the sound set increases. In two subsequent studies Caclin et al. [2006, 2007] found separate processing channels for the salient timbre dimensions (i.e., separate representations in the auditory sensory memory) and also evidence that there is a certain amount of crosstalk between these channels most probably occurring in later processing stages.

It also seems that the treatment of timbre as a merely continuous entity might be an additional cause of confusion. As mentioned above, McAdams et al. [1995] and McAdams [1999] suggested that timbre is a combination of continuous perceptual dimensions and discrete features (specificities) to which listeners are differentially sensitive. Lakatos' [2000] findings also supported the duality of timbre perception. He examined a combined sound set including both pitched and percussive instruments and concluded that timbre perception is both continuous and categorical. MDS revealed the continuous dimensions (spectral centroid and rise time) which were independent from musical training but cluster analysis indicated that sounds (especially percussive) were categorised based on source properties.

The descriptors that have exhibited the most significant correlations with perceptual dimensions according to the literature are discussed in the following subsections. Section 3.2.2 presents the full set of audio features examined for the purposes of this work along with their formulas.

2.4.1 Temporal envelope

One of the grey areas in timbre research is the salience of attack time. Previous classification studies have supported contradictory views regarding the perceptual significance of attack time

(i.e., attack time is perceptually more important, equally important or less important compared to the steady state). Hajda [2007] gives a concise overview of the literature and suggests that the inconsistent results regarding salience of time envelope characteristics are due to the lack of robust operational definitions based on signal characteristics. McAdams [1999] also points out that feature extraction algorithms make enormous computational errors for certain acoustically produced sounds, (according to this author's personal experience, this is particularly true for attack time estimation). Strong [in Luce, 1963, p.90] calculated two different attack times of non-percussive tones based merely on amplitude or merely on waveform structure of the signal. The *amplitude transient* was defined as the time required for the amplitude to reach 90% of the steady state amplitude and the *structure transient* was defined as the time required for the waveform to obtain the same structural characteristics as in the steady state. For most of the instruments the structure transient was found to be shorter than the amplitude transient showing that musical signals reach a steady structural state earlier than reaching their steady amplitude state. It should also be noted that defining the amplitude of the steady state might not always be a trivial task especially in the case where high amplitude modulations (tremolo) are present.

Hajda et al. [1997] proposed a model for segmenting the temporal envelope of continuant signals in perceptually relevant parts. The model was named *amplitude/centroid trajectory (ACT)* and partitioned the signal into 4 parts based on the relationship between the global spectral centroid and the root mean square (RMS) amplitude trajectories. The model made the assumption that, during the transient, the spectral centroid rises and falls abruptly (attack) and then gradually rises again together with the amplitude until they both reach their average level (attack/steady-state transition). The two trajectories then vary around their mean values (steady state) and finally they both rapidly decrease (decay). The efficacy of the model was tested through identification studies [Hajda, 1996, 1997, 1999] which revealed that it is not suitable for impulsive tones and that the identification of an instrument is a case dependent, complex process that can not be easily explained through a single rule. Caetano et al. [2010] have enhanced the perceptual efficiency of the ACT model by proposing an improved amplitude envelope estimation method that automatically detected the boundaries of the envelope regions. Their approach utilised a technique known as the true amplitude envelope (TAE) that optimally fits a curve by trying to match the peaks of the waveform.

Relational timbre studies that have included both continuant and impulsive sounds have

concluded that attack time is indeed associated with one of the timbre space dimensions [e.g. Krumhansl, 1989, McAdams et al., 1995, Kendall et al., 1999]. However, Iverson and Krumhansl [1993] supported that the perceptually salient dynamic characteristics of sounds are present throughout the tones since the similarity judgements on complete tones corresponded both to the judgements based just on the attack or the decay portions. Krimphoff [1993] and McAdams et al. [1995] correlated [$|r| = 0.94$, $p < 0.0001$ for both] the primary dimension of their MDS spaces with the logarithm of rise time:

$$LRT = log_{10}(t_{max} - t_{thresh}), \tag{2.5}$$

where t_{max} is the rise time from onset to maximum RMS amplitude and t_{thresh} is the time from onset until the amplitude is 2% of the amplitude at t_{max}. The technical report by Peeters [2004] proposed one additional method to estimate the beginning and end of the attack time. It is called the "*weakest effort method*" and calculates moving thresholds based on the behaviour of the signal during the attack.

2.4.2 Spectral envelope

Spectral shape can be divided into the distribution of energy and the the spectral fine structure.

Spectral energy distribution

Spectral energy distribution is the most characteristic aspect of sound quality and is mostly represented through the time-varying spectral centroid (SC) as mentioned above:

$$SC(t) = \frac{\sum_{n=1}^{N} f_n A_n(t)}{\sum_{n=1}^{N} A_n(t)}, \tag{2.6}$$

where f_n is the frequency and $A_n(t)$ is the amplitude of the *nth* partial of a spectrum with N frequency components at time t. Grey and Gordon [1978], Ehresman and Wessel [1978], Lakatos [2000], Kendall et al. [1999], McAdams et al. [1995] among others, have found very strong correlations between the mean of the SC and one of their MDS spaces dimension.

Spectral fine structure

Krimphoff [1993] has also examined the relationship of spectral fine structure with the third MDS dimension identified by Krumhansl. The first descriptor that he used was the time-varying *Odd*

Even Ratio (OER) which measures the ratio of the energy contained in odd harmonics versus the energy contained in even harmonics:

$$OER(t) = \frac{\sum_{n=2i-1}^{N/2} A_n^2(t)}{\sum_{n=2i}^{N/2} A_n^2(t)} \tag{2.7}$$

The second descriptor examined was initially labelled *Spectral Deviation* and afterwards renamed as *Spectral Irregularity* by Krimphoff et al. [1994] and McAdams et al. [1995] or *Spectral Smoothness* by McAdams et al. [1999]. This descriptor calculates the normalised sum of the deviation of each harmonic amplitude from the average of the three adjacent harmonic amplitudes (centred on the harmonic under study) and has yielded the highest correlation with the third Krumhansl perceptual dimension. The formula that calculates this descriptor is the following:

$$SI(t) = \frac{\sum_{n=2}^{N-1} \left| A_n(t) - \frac{A_{n+1}(t) + A_n(t) + A_{n-1}(t)}{3} \right|}{\sum_{n=1}^{N} A_n(t)}, \tag{2.8}$$

where the harmonic amplitudes $A_n(t)$ can be either logarithmic or linear.

2.4.3 Spectrotemporal characteristics

Another physical characteristic that has been linked with timbre perception is spectrotemporal variation (i.e., the amount of variation of the spectrum over time). Various researchers have come up with different metrics (also labelled differently) in order to measure the spectrotemporal characteristics of musical signals. The phenomenological observation that spectrotemporal behaviour is significant for timbre perception was made quite early on [e.g. Grey, 1977, Ehresman and Wessel, 1978] but was not quantified until the 1990s. Kendall and Carterette [1993b] captured spectral variation through a global descriptor called the *Mean Coefficient of Variation* (MCV):

$$MCV = \frac{\sum_{k=1}^{N} \frac{\sigma_n}{\mu_n}}{N}, \tag{2.9}$$

where σ_n is the standard deviation of the amplitude of frequency component n across time, μ_n is the mean amplitude of component n, and N is the number of frequency components analysed, in this case $N = 9$.

Krimphoff [1993], in his Masters thesis, introduced a number of acoustic correlates for musical timbre. Trying to associate the third dimension of Krumhansl's [1989] MDS space with some physical property, Krimphoff came up with three different global measures of spectral fluctuation. *Spectral variation* [Krimphoff et al., 1994] was defined as 1 minus the normalised cross-correlation between successive amplitude spectra $A_n(t-1)$ and $A_n(t)$:

$$spectralvariation = 1 - \frac{\sum_{n=1}^{N} A_n(t-1)A_n(t)}{\sqrt{\sum_{n=1}^{N} A_n(t-1)^2}\sqrt{\sum_{n=1}^{N} A_n(t)^2}} \tag{2.10}$$

Spectral flux was calculated as the mean deviation of the spectral centroid of each analysis window relative to the long term average spectral centroid. Finally, *Coherence* measured the synchronicity of the harmonics during the attack time.

2.5 Musical Timbre Semantics

The two previous sections discussed the formulation of perceptual timbre spaces and the salient acoustic correlates of their dimensions. This section will focus on the semantics of musical timbre and will present the techniques that have been applied for their investigation along with the most significant results.

Koelsch [2011] separated musical meaning into three different classes: *extra-musical*, *intra-musical* and *musicogenic* meaning. *Extra-musical* meaning refers to the interpretation of musical information in terms of extra-musical concepts. *Extra-musical* meaning is further divided into three subcategories: *iconic* (i.e., musical similes and metaphors), *indexical* (originating from action-related sound patterns indexing an intention or emotion) and *symbolic* musical meaning (referring to extra-musical associations due to conditioning from a certain culture)[4]. *Intra-musical* meaning emerges from interpreting structural units of music (e.g. harmonic sequence, rhythmical patterns, large scale structural relations etc.) and *musicogenic meaning* refers to personalised responses to musical stimuli (e.g. physical activity, emotional responses, self-related responses due to conditioning etc.). Musical timbre semantics generally fall into the subcategory of *iconic musical meaning*.

Another study on verbal description of timbre [Wake and Asahi, 1998] has essentially broken

[4]Koelsch based this division on similar subcategories introduced by Peirce [1931/1958] and first applied to music by Karbusicky [1986].

down what Koelsch calls *iconic musical meaning* into three subcategories of sound description: *sound itself* (e.g. onomatopoeia), *sounding situation* (e.g. sound source) and *sound impression* (e.g. adjectival description). Based on a visual information processing model suggested by Kawachi [1995], Wake argued that recognition and impression are independently and sequentially processed (i.e., sound impression can be perceived and described either independently or after recognition of the sound source). Wake concluded that even though sound impression description is the less frequent among the three, it can be particularly useful to represent sounds with unknown sound source. The rest of this section will cover literature on the lexical description of timbral impression.

Efficient as pairwise dissimilarity tests and MDS analysis may be for identifying perceptual timbre spaces, they are incapable of applying semantic labels to the dimensions. The labelling of the dimensions in such cases often comes as a result of some speculative interpretation. It is reasonable to assume that the mapping between a semantic and a perceptual timbre space must be complex and partial since not all perceivable attributes of sound can be adequately verbalised and also because verbalisation might be a product of conditioning. However, verbal description of sound quality and its association with physical properties of sound has intrigued researchers for a long time. von Helmholtz [1877, p. 118-119] has made one of the first systematic attempts to associate semantic attributes with acoustic characteristics and Lichte [1941] has broken down the timbre of complex tones into three independent semantic components, namely, *brightness*, *roughness* and *fullness*. Schaeffer [1966, p.232] has also noted that one can refer to "the timbre of a sound without attributing it to a given instrument, but rather in considering it as a proper characteristic of this sound, perceived per se" (cited by Donnadieu [2007, p. 272]). The technological advances of the past decades in the field of sound processing and sound synthesis have enabled practical applications of timbre semantics. Therefore, the potential development of a common, language-independent semantic framework for timbre description is highly desirable, as it could be exploited for the creation of intuitive sound synthesis and sound processing systems.

As a result, a complementary approach that aims to investigate semantics of timbre has been adopted by many researchers. The objective in this case is the elicitation of verbal descriptors, usually in the form of adjectives [von Bismarck, 1974a,b, Kendall and Carterette, 1993a,b]. According to this method, sound objects are represented by a feature vector of semantic attributes rather than by their relative perceptual distances. This is based on the hypothesis that timbre can

be adequately described by the use of semantic scales [Samoylenko et al., 1996]. The concept of using verbal attributes has also been applied for describing properties of specific musical instruments and characteristics of their performance [Disley and Howard, 2004, Nykänen et al., 2009, Barthet et al., 2010b, Fritz et al., 2012, Saitis et al., 2012, Traube et al., 2008], polyphonic timbre [Alluri and Toiviainen, 2010] and acoustic assessment of concert halls [Lokki et al., 2011]. An overview of various methods that can be used for elicitation of verbal descriptions is provided by Neher et al. [2006].

When the major objective is to investigate verbal description of musical timbre, then methods like *semantic differential* [Osgood et al., 1957 and Lichte, 1941, von Bismarck, 1974a] and one variant of this method, *verbal attribute magnitude estimation* (*VAME*) [Kendall and Carterette, 1993a,b] are usually employed instead of MDS. Whereas with the semantic differential each sound is rated along scales whose endpoints are labelled by two opposing verbal attributes such as 'bright-dull', with the VAME method the endpoints of the scales are labelled by an attribute and its negation ('not harsh-harsh'). These multidimensional data are then analysed by dimension reduction techniques such as Principal Components Analysis (PCA) [e.g. von Bismarck, 1974a, Kendall et al., 1999, Lokki et al., 2011] or Factor Analysis (FA) [e.g. Alluri and Toiviainen, 2010] and by Cluster Analysis techniques [e.g. Kendall and Carterette, 1993a, Disley et al., 2006] in order to achieve the reduction of a large number of semantic descriptions to a smaller number of interpretable factors.

One of the most cited studies on verbal description of timbre was conducted by von Bismarck [1974a,b] in German. He performed a semantic differential listening test featuring 30 verbal scales in order to rate the verbal attributes of 35 steady-state synthetic tones. The four dimensions identified by von Bismarck were labelled: *full-empty*, *dull-sharp*, *colourful-colourless* and *compact-diffused*[5]. Other related studies have also identified three or four semantic axes. Pratt and Doak [1976], working with simple synthetic tones and English adjectives, proposed a 3-D space featuring the dimensions: *bright-dull*, *warm-cold* and *rich-pure*. Štěpánek's study [2006] in Czech and German revealed the following dimensions for violin and pipe organ sounds: *gloomy-clear*, *harsh-delicate*, *full-narrow* and *noisy/rustle*. Moravec's work [2003], also in Czech, acquired descriptors through a questionnaire for timbre description without the presentation of any stimuli. It also resulted in the proposition of four semantic axes namely: *bright/clear-*

[5] '-' will be used to indicate antonyms and '/' will be used to indicate synonyms.

gloomy/dark, *hard/sharp-delicate/soft*, *wide-narrow* and *hot/hearty*. Finally, Disley's [2006] study in English used strings, brass, woodwind and percussive stimuli from the MUMS sound library [Opolko and Wapnick, 2006] and uncovered four salient dimensions labelled by the terms: *bright/thin/harsh-dull/warm/gentle*, *pure/percussive-nasal*, *metallic-wooden* and *evolving*.

The inhomogeneity observed in the above studies could be potentially attributed to factors related to method, stimuli or language. Štěpánek [2006] has proposed that semantic dimensions of timbre are dependent from pitch and instrument type, and Krumhansl and Iverson [1992] have also concluded that pitch and timbre are not perceived independently. This implies that the variety of stimuli and pitches used in the different studies could be responsible for the diversity in identified semantic dimensions. Furthermore, the data acquisition (selection and number of verbal descriptors) and analysis approaches (PCA, FA, etc.) also varied among the aforementioned studies. Finally, language is another potential factor of influence on timbre semantics. It has been argued that people's thinking about objects (including object description) is affected by grammatical differences across languages [Boroditsky et al., 2003]. Additionally, it has been reported that the use of some descriptive adjectives differs even between UK and US English speakers [Disley and Howard, 2004]. Therefore, more solid conclusions regarding the influence of language on semantic descriptions of timbre will require careful control of several factors.

2.6 Bridging semantics with perception

The previous sections have presented the typical methodology that is being followed when studying the perception and semantics of musical timbre. One major question that this thesis will try to address concerns the relationship between timbre perception and its semantics. The literature in the field is inconclusive, albeit there is evidence that semantic description conveys some meaningful perceptual information. Researchers have adopted various approaches to address this problem.

Kendall and Carterette [1993a,b] and Kendall et al. [1999] attempted to exploit a combination of pairwise dissimilarity and verbal attribute ratings for isolated and dyad timbres. The perceptual and semantic timbre spaces that resulted from these two approaches were compared but were found to be only partially similar. Faure et al. [1996] have also tried to bridge semantics with perception through a pairwise dissimilarity test and additional free verbal description of the perceptual distances. This study identified 22 semantic descriptors and associated them

with perceptual dimensions and acoustic characteristics. The majority of the adjectives correlated with more than one perceptual dimension. Therefore, the value of musical timbre description by verbal means remained an open question.

Other studies have also addressed this issue from different perspectives. From a linguistics perspective, Samoylenko et al. [1996] found that verbal description of perceived timbral dissimilarities corresponded well with numerical dissimilarity ratings. Therefore, a relationship between timbre description and timbre dissimilarity was suggested, but as stated by the authors, a remaining question was whether this relationship held up at the level of timbre space dimensions. The subsequent work of Kendall et al. [1999] found only weak support for the relationships requested by Samoylenko et al. [1996].

Furthermore, timbre semantics have recently been investigated through a neuroscientific approach which offered new insight to the question of meaning conveyed by timbre. Painter and Koelsch [2011] carried out two EEG experiments that demonstrated the ability of musical timbre to carry extra-musical meaning. More specifically, it has been demonstrated that prior listening to a sound can significantly influence the meaningful processing of a subsequent word or sound.

Alluri and Toiviainen [2010] have also identified three salient perceptual dimensions for polyphonic timbre, namely *activity*, *brightness* and *fullness*. In a subsequent study, Alluri et al. [2012] investigated the neural underpinnings of timbral and other features of a naturalistic musical stimulus. The acoustic parameters representing the basic perceptual timbre dimensions were identified and functioning Magnetic Resonance Imaging (fMRI) was utilised to localise parts of the brain that were responsible for processing each of these separate dimensions.

The above suggest that semantic description of musical timbre can provide significant information regarding perceptual representation of sound. However, this has not been adequately validated through comparison of pairwise dissimilarity rating (perceptual spaces) and verbal description studies (semantic spaces).

2.7 Interdependencies of timbre perception with pitch and auditory environment

2.7.1 Pitch

As previously stated, timbre has been studied mostly by trying to equalise for the other auditory attributes (i.e., pitch, subjective duration and loudness). However, not everyone has shared the opinion that pitch and timbre are two separate attributes of auditory sensation. Arnold Schoen-

berg, prominent composer and music theorist of the 20th century wrote:

> I cannot readily admit that there is such a difference, as is usually expressed, between timbre and pitch. It is my opinion that the sound becomes noticeable through its timbre and one of its dimensions is pitch. In other words: the larger realm is the timbre, whereas the pitch is one of the smaller provinces. The pitch is nothing but timbre measured in one direction. If it is possible to make compositional structures from timbres which differ according to height, [pitch] structures which we call melodies, sequences producing an effect similar to thought, then it must also be possible to create such sequences from the timbres of the other dimension from what we normally and simply call timbre. Such sequences would work with an inherent logic, equivalent to the kind of logic which is effective in the melodies based on pitch. All this seems a fantasy of the future, which it probably is. Yet I am firmly convinced that it can be realised. [Schoenberg, 1922, p.471]

The pitch-timbre interaction has been studied by various researches as also mentioned above [Krumhansl and Iverson, 1992, Štěpánek, 2006]. Miller and Carterette [1975] conducted a pairwise similarity experiment with tones of variable fundamental frequency (F_0) and identified pitch as a salient dimension of the perceptual space. Krumhansl and Iverson [1992] looked at the perceptual interactions between pitch and timbre working with both isolated tones and with longer sequences. They found that while pitch perception is robust to timbre variations the opposite does not hold true. This result suggested that patterns of timbre variation could not be easily attended unless pitch was held constant. Indeed, Handel and Erickson [2004] showed that pitch differences can confuse instrument identification, however, Vurma et al. [2010] supported that judgements of small pitch differences can also be affected by timbral variations. In other words, the dependency between timbre and pitch is bidirectional. Of course, this interdependency does not imply that it is not possible to compare timbres of different pitches. Marozeau et al. [2003], Marozeau and de Cheveignè [2007], for example, have shown that timbre differences are perceived independently from pitch differences at least within the range of one and a half octave. In a subsequent pairwise dissimilarity rating study, Marozeau and de Cheveignè [2007] found that auditory *brightness* (as predicted by the spectral centroid) is affected by F_0 a fact that was additionally supported by Schubert and Wolfe [2006] through a semantic description listening test.

Overall, the above findings demonstrate that timbre is influenced by pitch and vice versa. Handel and Erickson [2004] suggested that the independence shown in some cases could be the result of studying a big range of timbral variation compared to pitch differentiation and the opposite.

2.7.2 Auditory environment

In the real world, the listening of isolated sounds in the sterile vacuum of perfect silence is rare. Sounds (whether musical or not) usually exist in combination with other sounds and their interactions create complex soundscapes, ranging from a buzzing pub to a symphony orchestra performance. While timbre dependency on pitch has received considerable attention, the interdependency of simultaneously sounding timbres (i.e., polyphonic timbre) has been less studied.

Sandell [1995] has divided the concurrent presence of timbres into three categories, namely *timbral heterogeneity*, *timbral augmentation* and *emergent timbre*. *Timbral heterogeneity* describes the situation where two or more sound sources are concurrently active but are perceived as separate entities. Auditory Scene Analysis (ASA) theory [Bregman, 1994] uses the term *perceptual segregation* referring to the same phenomenon. Sound streams can be segregated based on timbral differences (e.g. different instruments playing in an ensemble) or merely based on contextual difference (e.g. two instruments of the same timbre class playing different melodies). Another example of perceptual segregation is the so called cocktail party effect, which describes the ability of a listener to focus on a single conversation in a noisy environment. *Timbral augmentation* and *emergent timbre* can be thought of as part of *perceptual fusion* which is the alternative option regarding concurrent sound streams offered by ASA theory. Fusion or blending occurs when two or more concurrent sounds are perceived as a single entity. *Timbral augmentation* refers to the special case where the timbre of a dominant sound is enhanced by the presence of another sound and *emerging timbre* describes the fusion of various timbral components into a novel percept. In both cases, the resulting overall timbre is a single percept.

An example of the few studies on more complex timbres is the work of Kendall and Carterette [1991, 1993a,b] on semantic description and pairwise dissimilarity judgements of wind instrument dyad tones. As also mentioned in section 2.6, Alluri and Toiviainen [2010] investigated polyphonic timbre perception. They showed that semantic ratings of polyphonic timbre are consistent across individuals and that the major semantic dimensions of polyphonic timbre, namely: activity, brightness and fullness, appear to be similar with the most commonly suggested seman-

tics of monophonic timbre. Subsequently, Alluri et al. [2012] utilising fMRI, investigated the areas of the brain that were activated by the previously identified timbral semantic dimensions of a naturalistic stimulus. They found a significant overlap among the brain areas that processed timbral dimensions (i.e., activity, brightness and fullness) but hardly any overlap between the areas responsible for processing timbral, rhythmic or tonal information.

The influence of background noise on timbre has received even less attention compared to polyphonic timbre despite the fact that music nowadays is very often being enjoyed in noisy environments (i.e., in means of transportation or in the street through MP3 players, in bars, in live gigs etc.). While the influence of the auditory environment on loudness has been investigated and modelled (e.g. Moore et al. [1997]), the same has not yet happened with timbre. Loudness relationships among spectral components of a single sound or of an auditory scene may affect timbre perception. As a result, it should be expected that if loudness is affected by the auditory environment this would in turn have an impact on timbre perception. Therefore, the effect of the auditory scene on timbre perception is a research path worth following.

2.8 Summary

This chapter has presented the basic literature in the field of timbre perception. We have discussed the difficulty to reach a satisfactory timbre definition, the most popular approaches for the study of timbre perception, the common acoustic correlates, timbral semantics and the interdependencies of timbre with the remaining auditory attributes. According to the main body of existing work, timbre perception is context dependent, influenced by pitch and by the auditory environment and timbral judgements also seem to exhibit significant nonlinearities. Overall, the amount of work on timbre perception that was carried out during the last few decades is significant but the complex and multidimensional nature of timbre leaves much scope for further investigation.

This thesis will contribute to three of the less explored areas of the existing literature. Firstly, it will test the universality of timbral semantics, that is the extent to which language of description affects the salient dimensions of a semantic timbre space. Secondly, it will investigate the relationship between the semantic and the perceptual timbre space (i.e., compare the semantic dimensions with the underlying perceptual dimensions). Finally, it will make an initial step on investigating the effect of the auditory environment on timbre perception.

Chapter 3

Methods

This chapter briefly introduces the methods employed in this work in order to facilitate comprehension of the following chapters by the non-expert reader. The first section (3.1) describes the basic statistic tools utilised for the purposes of this thesis and highlights their contribution to the various data analyses. The second section (3.2) concerns the audio feature extraction process. It starts by briefly presenting the Spectral Modeling Synthesis (SMS) platform which was the main input signal representation used. It subsequently presents the complete set of acoustic descriptors that were extracted from the sounds under study.

3.1 Statistical Techniques

As mentioned in the previous chapter, two common statistical techniques for analysing psychometric data in the field of timbre perception are Factor Analysis and Multidimensional Scaling analysis. Factor analysis is a dimension reduction technique that is mostly used to identify the latent semantic dimensions that exist within a large group of semantic variables. MDS on the other hand, exploits distance data (similarities or dissimilarities between pairs of sounds) to create a spatial configuration of the stimuli, i.e. identify the salient perceptual dimensions. These major techniques along with other tools that have been utilised for the purposes of this work (e.g. Cluster Analysis, CATPCA transformation) are presented below. The statistical algorithms and equations are according to the SPSS algorithms [IBM, 2011].

3.1.1 MDS algorithms

ALSCAL

ALSCAL [Young et al., 1978] performs metric or non-metric Multidimensional Scaling by using an Alternating Least Squares approach to scaling [Takane et al., 1977]. It offers several individual differences options (weighted scaling) where, apart from the representation of objects by points in a Euclidean space, each individual dissimilarity matrix is also represented by a vector of weights in an additional individual differences space. We have used ALSCAL to analyse the data presented in chapter 4. ALSCAL algorithm starts with an initial stimulus configuration. The distances are computed based on a weighted Euclidean model:

$$d_{ijk}^2 = \sum_{a=1}^{r} w_{ka}(x_{ia} - x_{ja})^2 \tag{3.1}$$

where r is the number of dimensions, w_{ka} is the weight for source (i.e participant) k on dimension a, and x_{ia} and x_{ja} are the coordinates of stimulus i and j on dimension a. The first set of distances are computed from an initial configuration and are then updated according to an iterative procedure. In the case of ordinal data, distances are transformed into disparities through Kruskal's least-squares monotonic transformation. The disparities, which are in the same rank order as the data and fit the distances as well as possible, are subsequently normalised. The optimisation process aims at minimising a measure of error called Young's S-Stress 1 or Tanake-Young-de Leeuw formula:

$$SStress(1) = \left[\frac{1}{m} \sum_{k=1}^{m} \left[\frac{\sum_i \sum_j (d_{ijk}^2 - d_{ijk}^{*2})^2}{\sum_i \sum_j d_{ijk}^{*4}} \right] \right]^{1/2} \tag{3.2}$$

where m is the number of sources, d_{ijk}^* are the normalised disparity values and d_{ijk} are the distances calculated from Equation 3.1. The current value of S-Stress 1 is compared to the value of S-Stress 1 from the previous iteration. If the improvement is less than a specified value, iteration stops and the output stage has been reached. If not, the program re-estimates the subject weights and the stimulus coordinates.

Squared correlation index (RSQ)

RSQ (R-Squared) is the squared correlation of the input distances with the scaled N-dimensional space distances using MDS coordinates. It reflects the proportion of variance of the input distance

data accounted for by the scaled data. The higher the value (to a maximum of 1) the better the fit.

PROXSCAL

PROXSCAL (Proximity Scaling) [Commandeur and Heiser, 1993] performs MDS of proximity data to find a least-squares representation of the objects in a low-dimensional space. Similarly to ALSCAL, PROXSCAL also offers metric and non-metric MDS, as well as options for weighted scaling for multiple dissimilarity matrices. However, ALSCAL uses the Young's S-Stress 1 formula for stopping its iterative solution procedure. This criterion can yield sub-optimal solutions [Coxon and Jones, 1980, Ramsay, 1988, Weinberg and Menil, 1993] as it attributes greater weights to larger dissimilarities, which are generally associated with greater error [Ramsay, 1988]. Thus, PROXSCAL is now generally preferred and it has been the favoured MDS analysis method for the data presented in chapters 6 and 7. PROXSCAL minimises the following loss function:

$$\sigma^2 = \frac{1}{m}\sum_{k=1}^{m}\sum_{i<j}^{n} w_{ijk}[\hat{d}_{ijk} - d_{ij}(\mathbf{X}_k)]^2 \tag{3.3}$$

which is the weighted mean squared error between the transformed proximities ($\hat{d}_{ijk}$) and the distances ($d_{ij}(\mathbf{X}_k)$). The number of objects (in our case sound stimuli) is n and the number of sources (in our case participants) is m. $\mathbf{X}_k$ is an $n \times p$ matrix (p = number of dimensions) with individual space coordinates for each source k. The distances $d_{ij}(\mathbf{X}_k)$ are the Euclidean distances between object points, with the coordinates in the rows of $\mathbf{X}_k$. The transformation function for the proximities provides nonnegative, monotonically nondecreasing values for the transformed proximities $\hat{d}_{ijk}$. w_{ijk} is a weight applied on each separate sound pair for each individual participant.

The PROXSCAL algorithm consists of four major steps:

1. find initial configuration $\mathbf{X}_k$ and evaluate the loss function;
2. find an update for the configurations $\mathbf{X}_k$;
3. find an update for the transformed proximities $\hat{d}_{ijk}$;
4. evaluate the loss function; if some of the predefined stop criterion is satisfied, stop; otherwise, go to step 2.

S-Stress

S-Stress[1] is a measure of misfit given by Equation 3.4. It measures the difference between inter-point distances in computed MDS space and the corresponding actual input distances. The lower the value (to a minimum of 0) the better the fit.

$$SStress = \eta^4(\hat{\mathbf{D}}) + \eta^4(\alpha\mathbf{X}) - 2\rho^2(\alpha\mathbf{X}) \tag{3.4}$$

where

$$\alpha^2 = \frac{\rho^2(\mathbf{X})}{\eta^2(\mathbf{X})} \tag{3.5}$$

$$\eta^4(\hat{\mathbf{D}}) = \sum_{k=1}^{m}\sum_{i<j}^{n} w_{ijk}\hat{d}_{ijk}^4 \tag{3.6}$$

$$\eta^2(\mathbf{X}) = \sum_{k=1}^{m}\sum_{i<j}^{n} w_{ijk}d_{ij}^2(\mathbf{X}_k) \tag{3.7}$$

$$\eta^4(\mathbf{X}) = \sum_{k=1}^{m}\sum_{i<j}^{n} w_{ijk}d_{ij}^4(\mathbf{X}_k) \tag{3.8}$$

$$\rho^2(\mathbf{X}) = \sum_{k=1}^{m}\sum_{i<j}^{n} w_{ijk}\hat{d}_{ijk}^2 d_{ij}^2(\mathbf{X}_k) \tag{3.9}$$

Dispersion Accounted For (DAF)

DAF is a measure of fit. The higher the value (to a maximum of 1) the better the fit. It is calculated by the following equation:

$$DAF = 1 - \sigma^2 \tag{3.10}$$

where σ^2 is calculated from Eq. 3.3.

3.1.2 Hierarchical Cluster Analysis

Cluster Analysis is a statistical technique that seeks to identify homogeneous subgroups of variables (or cases) within a larger set of observations [Romesburg, 2004]. Hierarchical clustering

[1]Note that the Young's S-Stress 1 of the ALSCAL algorithm mentioned above is not the same metric as the PROXSCAL S-Stress.

is one of the available clustering algorithms that starts with each variable (or case) in a separate cluster and combines clusters until only one is left. As will be discussed in chapters 5 and 7 hierarchical clustering has been used both to indicate groups of semantically related verbal descriptors and to examine the structure of a timbre space (where the observations are the positions of the sound stimuli within the perceptual space).

1. Begin with N clusters each containing one variable.

2. Find the most similar pair of clusters p and q ($p > q$) and denote the dissimilarity or similarity as s_{pq}. The distance measures vary, e.g. simple Euclidean, squared Euclidean, Pearson correlation, Chebychev, Minkowski.

3. Merge clusters p and q into a new cluster $t(=q)$ and update the dissimilarity or similarity matrix S (by the specified distance measure) to represent revised dissimilarities or similarities (s_{tr}) between cluster t and all other clusters r. Delete the row and column of S corresponding to cluster p.

   i) The formula that calculates s_{tr} for the centroid linkage method (employed in chapter 5) is the following:

$$s_{tr} = \frac{N_p}{N_p + N_q} s_{pr} + \frac{N_q}{N_p + N_q} s_{qr} - \frac{N_p N_q}{(N_p + N_q)^2} s_{pq} \tag{3.11}$$

   where N_i represents the number of variables (or cases) in cluster i.

   ii) and the formulas for the average linkage (employed in chapter 7) is:

$$s_{tr} = s_{pr} + s_{qr} \tag{3.12}$$

   and

$$N_t = N_p + N_q \tag{3.13}$$

   The most similar pairs are then chosen based on the value

$$s_{ij}/(N_i N_j) \tag{3.14}$$

4. Perform steps 2 and 3 until all entities are in one cluster.

Both the centroid and the average linkage methods are the hierarchical clustering methods that are less affected by outliers since the first compares cluster means and the later considers all members in the cluster.

Average Silhouette Width Validity Index (ASWVI)

ASWVI is a cluster evaluation metric. The Silhouette Width Validity Index of the point i is given by:

$$SWVI = \frac{min\{D_{ij}, j \in C_{-i}\} - D_{ic_i}}{max(min\{D_{ij}, j \in C_{-i}\}, D_{ic_i})} \tag{3.15}$$

where C_{-i} represents clusters that do not include point i as a member, c_i is the cluster which includes point i and D_{ij} is the distance between point i and the centroid of cluster j. If the denominator equals zero, the SWVI of point i is not included in calculation of the average SWVI from the following equation.

$$ASWVI = \frac{1}{N}\sum_{i=1}^{N} \frac{min\{D_{ij}, j \in C_{-i}\} - D_{ic_i}}{max(min\{D_{ij}, j \in C_{-i}\}, D_{ic_i})} \tag{3.16}$$

ASWVI can take values from -1 to 1. Values closer to 1 indicate a better assignment of points to clusters. In our case, ASWI was applied to evaluate the clustering of semantic variables (see section 5.3.2) before and after optimal transformation.

3.1.3 Factor Analysis

Factor Analysis (FA) [Harman, 1976] is a dimension reduction technique that aims at a parsimonious conceptual understanding of a group of measured variables. To this end, it determines the number, nature and relationships of some common factors in a way that better accounts for the pattern of correlations between the variables. As noted in the beginning of this chapter, FA is most appropriate when a researcher seeks to identify the underlying structure of a set of variables. The basic FA model is described as:

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jn}F_n + U_j = \sum_{i=1}^{n} a_{ji}F_i + U_j \tag{3.17}$$

where $j = 1 \ldots m$ or in matrix notation,

$$\mathbf{Z = A \cdot F + U} \tag{3.18}$$

where

$$\mathbf{Z}^T = \begin{bmatrix} z_1 & \cdots & z_m \end{bmatrix} \tag{3.19}$$

is the array of m analysed variables,

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \tag{3.20}$$

is the matrix of *factor loadings* to be estimated from the data,

$$\mathbf{F}^T = \begin{bmatrix} F_1 & \cdots & F_n \end{bmatrix} \tag{3.21}$$

is the array of *n Common Factors*, and

$$\mathbf{U}^T = \begin{bmatrix} U_1 & \cdots & U_m \end{bmatrix} \tag{3.22}$$

is the array of *m Unique Factors*.

As shown in the above equations, FA takes a set of original variables and creates a new set of constructs (the common factors, with n $<$ m) that will compactly describe the correlations between the original variables. Unique factors add to the versatility of the solution, as they account for that part of the original variance that cannot be attributed or modelled by the common factors. The frequently used PCA only achieves data reduction through maximization of the variance explained by the principal components and does not account for unique variance. That is, PCA determines the linear combinations of the variables under study (the principal components) that retain as much information from the original variables as possible. As a result principal components are not latent variables and should not be confused with common factors. Therefore, FA was deemed to be more appropriate for the exploratory study described in chapter 5 [Fabrigar et al., 1999].

FA methods

Two of the more commonly used FA algorithms are Maximum Likelihood (ML) and Principal Axis Factoring (PAF). ML allows for generalisation of the results beyond the particular sample under study. On the downside, ML requires multivariate normality among the variables. PAF, on the other hand, does not pose any distributional assumptions but its results should not be generalised. Aiming at generalisation of our findings, ML has been the favoured FA method for this work and as explained in subsection 5.3.2, a transformation applied on our data (see subsection 3.1.4) improved the conditions for its application.

3.1.4 CATPCA

CATegorical PCA (CATPCA) originally targeted the problem of including categorical variables in the analysis with numerical variables. The basic idea was the assignment of numerical quantifications (based on an optimising criterion) to the categories of each variable, thus allowing standard procedures to be used on the quantified variables. The categorisation of the variables is done automatically by grouping the values into categories with a close to 'normal' distribution. An iterative method called Alternating Least Squares (ALS) [De Leeuw et al., 1976] calculates the quantifications corresponding to each category which are then used to obtain a solution. The solution is subsequently used to update the quantifications which in turn produce a new solution until some criterion is satisfied. The optimising criterion for variables quantification aims at increasing the correlations between the object scores (scores of each object on each dimension) and each of the quantified variables, i.e. maximization of the reproduced variance. Applied to a numerical (as in our case) data matrix

$$\mathbf{D} = \begin{bmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nm} \end{bmatrix} \tag{3.23}$$

with n observations (i.e. objects) and m original variables (similarly to matrix $\mathbf{Z}$ and $z_1 \cdots z_m$ defined in subsection 3.1.3), such an optimisation criterion is equivalent to the minimisation of the following cost function:

$$\sum_{k=1}^{n} \sum_{j=1}^{m} (\tilde{d}_{kj} - \mathbf{x_k a_j^T})^2 \tag{3.24}$$

or in matrix notation

$$L(\tilde{\mathbf{D}}, \mathbf{X}, \mathbf{A}) = \sum_{j=1}^{m} \|\tilde{\mathbf{D}}_{\mathbf{j}} - (\mathbf{X}\mathbf{A}^{\mathbf{T}})_{\mathbf{j}}\|^2 \tag{3.25}$$

where $\tilde{\mathbf{D}}_{\mathbf{j}}$ is the j_{th} column (i.e. variable) of the $(n \times m)$ matrix $\tilde{\mathbf{D}} = \varphi(\mathbf{D})$ of optimally transformed data, φ is the set of nonlinear transformations of the original variables (columns of the original data matrix $\mathbf{D}$), $\mathbf{a}_{\mathbf{j}}$ is the j_{th} row of the $(m \times p)$ matrix $\mathbf{A}$ of component loadings (as defined in 3.1.3), p is the number of selected principal components and x_k is the k_{th} row of the $(n \times p)$ matrix $\mathbf{X} = \begin{bmatrix} \mathbf{x_1} & \dots & \mathbf{x_n} \end{bmatrix}^T$ of objects scores in the component space. As previously mentioned, the minimisation of the above function over the possible nonlinear transformations of the original variables is performed in an iterative way, by alternating solutions based on object scores, component loadings and variables transformations, until a convergence criterion is satisfied. The obtained solution will also depend on the selection of the number of principal components, p. In the experiment that is described in chapter 5 our observations are the assessed sounds and our variables are the semantic descriptors.

CATPCA is also valuable for optimal nonlinear transformation of numerical variables. It should be noted that CATPCA does not assume linear relationships among numeric data nor does it require multivariate normal data. An additional important property of CATPCA is the fact that it allows for variables to be scaled at different levels of measurement namely: nominal, ordinal, monotonic and non-monotonic splines [Meulman and Heiser, 2008].

The CATPCA optimal transformation was applied to the semantic variables of a verbal attribute magnitude estimation (VAME) experiment presented in chapter 5. This was to better account for possible nonlinear relationships between variables that would otherwise be ignored by a simple factor analytic approach. Chapter 5 explains how these transformations have indeed contributed to a better modelling of our variable set.

3.1.5 Cronbach's Alpha

Cronbach's Alpha is a measure of internal consistency and is commonly used as an estimate of the reliability of a psychometric test. Reliability is mathematically defined as the proportion of the variability in the responses to a survey that can be attributed to differences between respondents rather than to poor design of the experiment. That is, differences in the collected data are caused by differences in the opinion or perception of the respondents rather than by confusing questions

or multiple interpretations. The Cronbach's Alpha is calculated by the following equation:

$$\alpha = \frac{k(\overline{cov}/\overline{var})}{1+(k-1)(\overline{cov}/\overline{var})} \tag{3.26}$$

where k is the number of items under study, $\overline{var}$ is the average item variance (the average of the diagonal of the variance-covariance matrix) and $\overline{cov}$ is the average inter-item covariance (the average of the off-diagonal elements of the variance-covariance matrix).

In the context of this work, Cronbach's Alpha has been used to test the consistency of pairwise dissimilarity judgements made by a group of listeners over a set of sound stimuli. In this case, k was the number of participants and the variables of the variance-covariance matrix were the vectors of the pairwise dissimilarity judgements. A commonly used rule of thumb is that $a \geq 0.7$ is regarded acceptable for cognitive tests [Kline, 1999, George and Mallery, 2003, p.231]. However such general interpretations should be used having in mind that Cronbach's Alpha depends on the number of items in the study and tends to increase when their number increases.

3.2 Acoustic descriptors and their computational extraction

As explained in chapter 2, one major objective of timbre perception studies is to associate perceptual dimensions with physical properties of sound. The physical properties are usually represented by a number of signal characteristics both in the time and in the frequency domain and a variety of input signal representations have been adopted by researchers for analysing audio signals. This work has mostly used the output of the Spectral Modeling Synthesis (SMS) model [Amatriain et al., 2002] as an input signal representation for the extraction of harmonic acoustic descriptors. Two other Matlab toolboxes, the MIR Toolbox [Lartillot et al., 2008] and the Timbre Toolbox [Peeters et al., 2011]), that compute most of our timbre descriptors have also been used. However, a comparison showed inconsistencies in the calculation of several descriptors. Thus, the calculation of the acoustic descriptors was made based on formulas from Peeters [2004] and Peeters et al. [2011] that were implemented through the SMS platform.

3.2.1 Spectral Modeling Synthesis

SMS was first introduced by Serra and Smith [1990] and models sounds based on the deterministic plus stochastic model, i.e., as a number of sinusoids (partials) plus noise (residual component). This representation imitates the sound production of musical instruments. For example, a pitched

sound presupposes the existence of a periodic vibration which corresponds to the deterministic part and any other sound that can not be accounted for by this vibration (e.g. bow noise, noisy transients, noise of a plucked string, breath noise) is modelled by the stochastic part.

Equation 3.27 shows the mathematical representation of the model.

$$s(t) = \sum_{n=1}^{N} A_n(t) cos[\theta_n(t)] + e(t) \tag{3.27}$$

where $A_n(t)$ and $\theta_n(t)$ are the instantaneous amplitude and phase of the *nth* sinusoid, respectively, and $e(t)$ is the noise component at time t in seconds.

During the analysis stage, the time-varying partials of a sound are detected and are then represented by time-varying sinusoids. These sinusoids are added to create the harmonic part of the signal which is subsequently subtracted from the original sound leaving only the noise or 'residual' part. The residual is modelled through a time-varying filtered white noise component as shown in Equation 3.28.

$$e(t) = \int_{0}^{t} h(t,\tau) u(\tau) d\tau \tag{3.28}$$

where $u(\tau)$ is white noise and $h(t,\tau)$ is the impulse response of a time-varying filter at time t. In other words, the residual part is modelled by the convolution of a time-varying frequency-shaping filter with white noise. The synthesis stage combines additive synthesis for the sinusoidal part and subtractive synthesis for the noise part. A detailed description of the SMS algorithm can be found in Serra [1997].

Heuristic tests of the algorithm parameters suggested that a window of 4096 samples ($f_s =$ 44.1 kHz) was suitable for analysis of the full range of our stimuli set (including both continuant and impulsive sounds) as it offered both the spectral and temporal precision required for an accurate re-synthesis of the signals. The hop size was set to 512 samples and the zero padding factor to 2. Fifty partials were extracted for all sounds. As Serra and Smith [1990] point out, SMS is problematic when it comes to sounds that include noisy partials. This was indeed the case with some of our most noisy sound stimuli (e.g. tenor saxophone, Acid and Moog synthesisers) where the algorithm had problems identifying the correct pitch and as a result separated the deterministic from the stochastic part inaccurately. A manual assignment of the fundamental frequency resolved this issue and allowed accurate re-synthesis of these sounds.

3.2.2 Formulas of acoustic descriptors

This section will present the formulas of all the acoustic descriptors that were extracted for the purpose of this work.

Temporal characteristics

- Attack time calculation

  Three methods for calculating attack time were employed in this work. The first one was based on the 'Amplitude/Centroid Trajectory' approach proposed by Hajda et al. [1997] and calculated the time needed for the spectral centroid to reach its first minimum. The second one was based on the approach by Zaunschirm et al. [2012] and calculated the duration of the attack by applying an adaptive threshold (median filter) and using *Spectral Flux* [Peeters, 2004] as the detection function. The third method used an adaptive threshold as described by Peeters [2004] in order to calculate the rise time of the energy envelope. These three attack times and their logarithms (Equations 3.29 and 3.30) were considered as possible acoustic correlates.

$$At\_time = t_{end} - t_{start} \tag{3.29}$$

$$Log\_At\_time = log_{10}(t_{end} - t_{start}) \tag{3.30}$$

- Temporal centroid

  Temporal centroid [Peeters et al., 2000] is the center of gravity of the root-mean-square (RMS) energy envelope $e(t)$. It distinguishes percussive from sustained sounds.

$$TC = \frac{\sum_t te(t)}{\sum_t e(t)} \tag{3.31}$$

  where

$$e(t) = \sqrt{\frac{1}{T}\sum_{i=1}^{T} x_i^2(t)} \tag{3.32}$$

  and T is the window length in number of samples, t is the hop size expressed in seconds and $x_i(t)$ is the i_{th} amplitude sample of the window centred around t.

- Zero Crossing Rate

  Zero crossing rate is a measure of the number of times the signal value $s(t)$ crosses the zero axis. The noisier the signal, the larger the value for a fixed amount of time. The computation of this feature takes place directly on the signal $s(t)$, where the local DC offset of each frame is first subtracted and subsequently the zero-crossings rate value of each frame is normalised by the window length in seconds.

Spectral shape

Statistical moments of the spectrum

Spectral centroid, *spectral spread*, *spectral skewness* and *spectral kurtosis* constitute the first four statistical moments of the spectrum.

- Harmonic Spectral Centroid

$$SC(t) = \frac{\sum_{n=1}^{N} f_n(t)A_n(t)}{\sum_{n=1}^{N} A_n(t)}, \tag{3.33}$$

  where $A_n(t)$ and $f_n(t)$ are the magnitude and frequency of the *nth* harmonic at time t respectively and N indicates the maximum number of harmonics taken into account. The harmonic spectral centroid is the barycentre of the harmonic spectrum.

- Normalised Harmonic Spectral Centroid

$$SC_{norm}(t) = \frac{\sum_{n=1}^{N} nA_n(t)}{\sum_{n=1}^{N} A_n(t)} \tag{3.34}$$

  The *normalised harmonic spectral centroid* is expressed in number of harmonics.

- Normalised Energy Harmonic Spectral Centroid

$$SC_{energy}(t) = \frac{\sum_{n=1}^{N} nA_n^2(t)}{\sum_{n=1}^{N} A_n^2(t)} \tag{3.35}$$

- Corrected Spectral Centroid

  A modified version of the SC in order to account for the effect of F_0 on auditory brightness was also estimated according to Marozeau and de Cheveignè [2007]. The calculation followed the steps below:

  1. Calculate the SC using Moore's instantaneous specific loudness [Moore et al., 1997] as signal representation.
  2. Convert the SC ERB-rate value to the corresponding value in Hz according to the formula: $\overline{f} = \frac{(exp(\overline{Z}/9.26)-1)}{0.00437}$, where $\overline{Z}$ is the ERB-rate value.
  3. Subtract F_0 from the SC in Hz: $\overline{f_{corrected}} = \overline{f} - F_0$.
  4. Re-convert $\overline{f_{corrected}}$ to ERB-rate: $SC\_loud\_cor = 9.26ln(0.00437\overline{f_{corrected}}+1)$

- Harmonic Spectral Spread or Spectral Standard Deviation

$$Spread^2(t) = \frac{\sum_{n=1}^{N}(n-SC(t))^2 A_n(t)}{\sum_{n=1}^{N} A_n(t)} \tag{3.36}$$

  Harmonic spectral spread represents the spread of the harmonic spectrum around its mean value.

- Harmonic Spectral Skewness

$$Skewness(t) = \frac{\sum_{n=1}^{N}(n-SC(t))^3 A_n(t)}{spread(t)^3 \sum_{n=1}^{N} A_n(t)} \tag{3.37}$$

  Harmonic spectral skewness gives a measure of asymmetry of the harmonic spectrum around its mean value. As shown in Figure 3.1, skewness = 0 indicates a symmetric distribution, skewness > 0 more energy on the left and skewness < 0 more energy on the right.

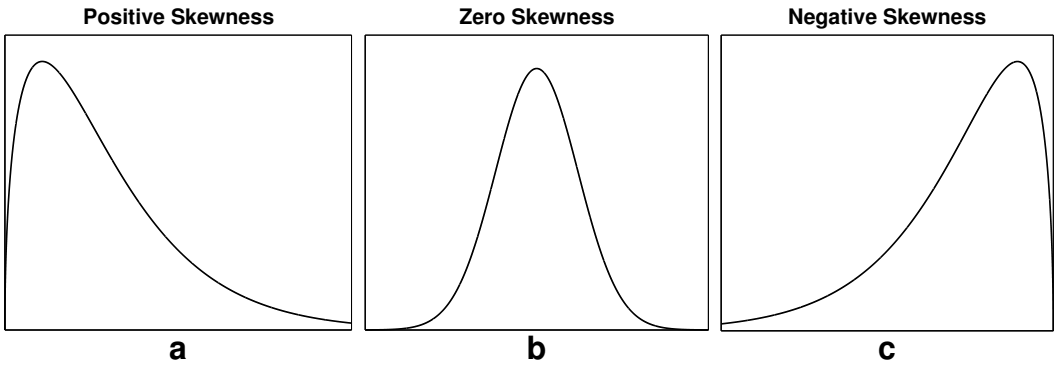


Figure 3.1: (a) Positive skewness, (b) zero skewness and (c) negative skewness.

- Harmonic Spectral Kurtosis

$$Kurtosis(t) = \frac{\sum_{n=1}^{N} (n - SC(t))^4 A_n(t)}{spread(t)^4 \sum_{n=1}^{N} A_n(t)} \tag{3.38}$$

Harmonic spectral kurtosis measures the flatness of the distribution around its mean value with kurtosis = 3 indicating a normal distribution, kurtosis $<$ 3 a flatter distribution and kurtosis $>$ 3 a peakier distribution, as shown in Figure 3.2.

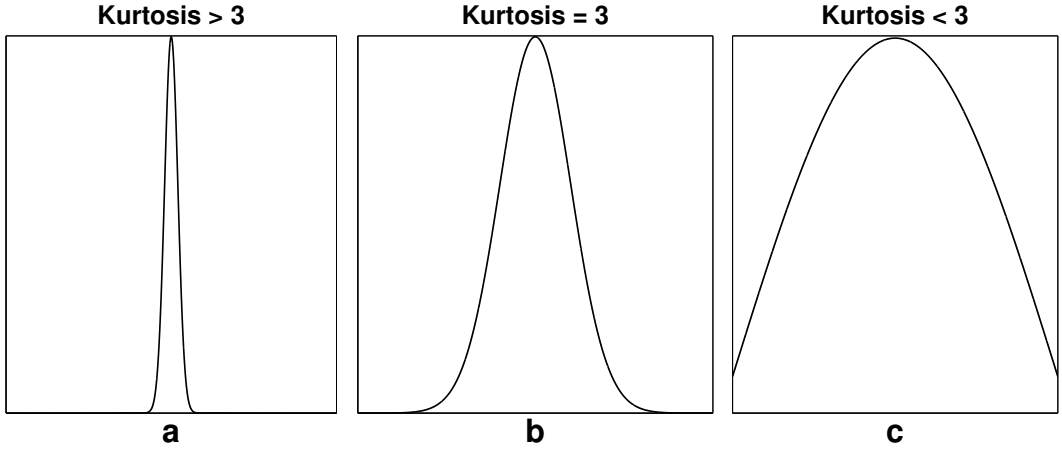


Figure 3.2: (a) Peaky distribution, (b) normally distributed (K=3), (c) flatter distribution.

Other spectral shape descriptors

- Harmonic Spectral Slope

$$a(f,t) = slope(t)f + const \tag{3.39}$$

where

$$Slope(t) = \frac{1}{\sum_{n=1}^{N} A_n(t)} \frac{N\sum_{n=1}^{N} f_n(t)A_n(t) - \sum_{n=1}^{N} f_n(t)\sum_{n=1}^{N} A_n(t)}{N\sum_{n=1}^{N} f_n^2(t) - (\sum_{n=1}^{N} f_n(t))^2} \tag{3.40}$$

Harmonic spectral slope represents the amount of decreasing of the spectral amplitude and is computed by linear regression of the spectral amplitude.

- Harmonic Spectral Decrease

$$Decrease(t) = \frac{1}{\sum_{n=2}^{N} A_n(t)} \sum_{n=2}^{N} \frac{A_n(t) - A_1(t)}{n-1} \tag{3.41}$$

Harmonic spectral decrease [Krimphoff, 1993] also represents the amount of decreasing of the spectral amplitude but it emphasises the slope of the lower frequencies.

- 95% Spectral Roll-Off

$$\sum_{f=0}^{f_c(t)} A_f^2(t) = 0.95 \sum_{f=0}^{SR/2} A_f^2(t) \tag{3.42}$$

where A_f are magnitudes of the f frequency bins, $f_c(n)$ is the *spectral roll-off* frequency for a particular frame and SR/2 is the Nyquist frequency. *95% Spectral roll-off* [Scheirer and Slaney, 1997] is the frequency below which 95% of the signal energy is contained.

Spectral fine structure

- Odd to Even Harmonic Energy Ratio

$$OER(t) = \frac{\sum_{n=2i-1}^{N/2} A_n^2(t)}{\sum_{n=2i}^{N/2} A_n^2(t)} \tag{3.43}$$

where the harmonic amplitudes $A_n(t)$ were calculated as both logarithmic and linear. Since the logarithmic OER did not feature any significant correlation with any of the perceptual

dimensions of our listening experiments, wherever OER is mentioned in the text it will actually stand for linear OER.

- Spectral Irregularity

$$SI(t) = \frac{\sum_{n=2}^{N-1} \left| A_n(t) - \frac{A_{n+1}(t) + A_n(t) + A_{n-1}(t)}{3} \right|}{\sum_{n=1}^{N} A_n(t)}, \tag{3.44}$$

where the harmonic amplitudes $A_n(t)$ were calculated as both logarithmic and linear. Spectral irregularity is the sum of deviations of each harmonic amplitude from the mean of three consecutive harmonic amplitudes (centred on that harmonic), normalised by a global mean amplitude.

Harmonic analysis

- Inharmonicity

$$Inharmonicity(t) = \frac{2}{f_0} \frac{\sum_{n=1}^{N} | f(n) - nf_0 | N^2(t)}{\sum_{n=1}^{N} A_n^2(t)} \tag{3.45}$$

Inharmonicity represents the divergence of the signal harmonic series from a purely harmonic signal. It ranges from 0 (purely harmonic signal) to 1 (inharmonic signal).

- Relative energy of the first three harmonics

$$W(t) = \frac{\sum_{n=1}^{3} A_n^2(t)}{\sum_{n=4}^{N} A_n^2(t)} \tag{3.46}$$

This descriptor has been proposed as the acoustic correlates for auditory *warmth* [Williams and Brookes, 2010].

- Tristimulus 1, 2 and 3

$$T_1(t) = \frac{A_1(t)}{\sum_{n=1}^{N} A_n(t)} \tag{3.47}$$

$$T_2(t) = \frac{\sum_{n=2}^{4} A_n(t)}{\sum_{n=1}^{N} A_n(t)} \tag{3.48}$$

$$T_3(t) = \frac{\sum_{n=5}^{N} A_n(t)}{\sum_{n=1}^{N} A_n(t)} \tag{3.49}$$

The tristimulus values are three different types of amplitude ratios that have been introduced by Pollard and Jansson [1982] as a timbre equivalent to the colour attributes in vision.

- Noisiness

$$noisiness(t) = \frac{E_N(t)}{E_T(t)} = \frac{E_T(t) - \sum_{n=1}^{N} A^2(t)}{E_T(t)} \tag{3.50}$$

Noisiness is the ratio of the noise energy ($E_{total} - E_{harmonic}$) to the total energy of the signal.

Spectrotemporal characteristics

- Mean Coefficient of Variation (MCV)

$$MCV = \frac{\sum_{n=1}^{N} \frac{\sigma_n}{\mu_n}}{N} \tag{3.51}$$

This feature was proposed by Kendall and Carterette [1993b] as an alternative of *spectral flux*. σ_n is the standard deviation of the amplitude of frequency component n across time, μ_n is the mean amplitude of component n, and N is the number of frequency components analysed, in this case $N = 9$.

- Harmonic Spectral Variation or Spectral Flux

$$spectralvariation = 1 - \frac{\sum_{n=1}^{N} A_n(t-1)A_n(t)}{\sqrt{\sum_{n=1}^{N} A_n(t-1)^2}\sqrt{\sum_{n=1}^{N} A_n(t)^2}} \tag{3.52}$$

Spectral variation represents the amount of variation of the spectrum along time defined

as 1 minus the normalised correlation between successive A_n. Spectral variation is close to 0 if successive spectra are similar, or close to 1 if successive spectra are very different.

For all of the above descriptors, in addition to their harmonic version, we calculated (whenever possible) the equivalent values using the FFT bins as input. Some specific features (e.g. spectral centroid, spectral spread, spectral variation) were also computed using the instantaneous specific loudness per ERB band as calculated by Moore's loudness model [Moore et al., 1997]. Finally, the mean, median, standard deviation, range, skewness and kurtosis of each descriptor were additionally computed in an effort to capture elements of the time variant behaviour of the sounds.

3.3 Summary

The first section of this chapter presented the basic statistic methods utilised in this work. Two MDS algorithms (ALSCAL and PROXSCAL) were described and the use of PROXSCAL for analysing dissimilarity matrices for the main part of this work was justified. Two algorithms of hierarchical cluster analysis (centroid and average linkage) were also presented and the use of cluster analysis within the context of this work was introduced. Factor Analysis was subsequently presented and its appropriateness over the most commonly used PCA for identifying the latent structure of a set of variables was highlighted. Finally, CATPCA technique and its optimal nonlinear transformation that was applied to the semantic variables of chapter 5 were discussed.

The second section was devoted to acoustic descriptors and their extraction process. It began with a short presentation of the SMS platform, whose harmonic analysis output was used as an input signal representation for feature extraction, and was completed by presenting the formulas of the acoustic descriptors used.

Chapter 4

Exploring the relationship between auditory brightness and warmth: a study on synthesised stimuli

---

4.1 Introduction

This chapter constitutes an initial piece of work on timbre semantics and their acoustic correlates. It is a narrowly focused approach which attempted to validate findings from a previous study [Williams and Brookes, 2010] that had associated auditory *brightness* and *warmth* with separate, but nevertheless related, audio descriptors. *Brightness* is arguably the most popular semantic descriptor of musical timbre and a number of studies have shown its high positive correlation with spectral centroid [e.g. Lichte, 1941, von Bismarck, 1974a, McAdams et al., 1995, Schubert et al., 2004, Williams and Brookes, 2007]. *Warmth* on the other hand, does not feature such a commonly acceptable acoustic correlate and some studies have shown a lesser or greater amount of overlap between *warmth* and *brightness* [Howard et al., 2007, Ethington and Punch, 1994, Pratt and Doak, 1976, e.g.]. Recent work by Williams and Brookes [2010] has proposed a timbre morphing technique for achieving independent *brightness-warmth* modification using the SMS (spectral modeling synthesis) platform [Serra and Smith, 1990]. In this work the acoustic correlate of *warmth* was defined as the relative percentage of energy in the first three harmonic partials (see Eq. 3.46).

Based on the above, we assumed that it might be possible to achieve independent modification of auditory *brigthness* and *warmth* by manipulating the spectral centroid independently from the relative energy of the first three harmonic partials. Therefore, a specific experiment that

aimed to examine the *brightness-warmth* relationship by testing this hypothesis was designed and conducted.

A two-part additive synthesis algorithm that could modify the normalised harmonic energy spectral centroid (see Eq. 3.35) independently from the acoustic correlate for *warmth* (Eq. 3.46) and vice versa was created. A set of two pairwise dissimilarity rating listening experiments, featuring stimuli synthesised by this algorithm, was subsequently conducted to evaluate the perceptual effect of this manipulation. The influence of the fundamental frequency on the results was additionally examined by testing two separate groups of identical stimuli differing only in fundamental frequency (at 220 and 440 Hz). Multidimensional Scaling (MDS) analysis applied to the results constructed 2-D spaces that revealed the perceptual relationships among the stimuli. The test was completed with a verbal elicitation part which aimed at applying semantic labels to the identified perceptual dimensions.

4.2 Additive synthesis

Additive synthesis was among the first synthesis techniques in computer music. Its first extensive description was made by Moorer [1977]. The method is based on the Fourier's theorem principle that any periodic signal may be modelled as the sum of a number of sinusoids with time-varying parameters, also called partials (or harmonics when they are harmonically related). Thus, additive synthesis produces sounds by adding a number of sine wave oscillators. It has been the preferred synthesis method for this task as it provides the highest level of control among all other sound synthesis methods. The mathematical representation of additive synthesis is shown in the following formula:

$$s(t) = \sum_{n=1}^{n_{max}} A_n(t)\cos(2\pi f_n(t)t + \phi_n) \tag{4.1}$$

where t is time, n is the number of the harmonic partial, $A_n(t)$ is the time varying amplitude of the *nth* harmonic partial, $f_n(t)$ is the time varying frequency of the *nth* partial and ϕ_n is the phase of the *nth* partial. Equation 4.1 is used to define the value of the time-domain waveform $s(t)$ at time t. Each of the parameters is continually evolving. Successive frequency and amplitude values are used to describe the evolution of each sinusoid, the summation of which can create complex wave shapes and rich timbres. Figure 4.1 demonstrates the effect of harmonic addition in the time-domain waveform of a signal.

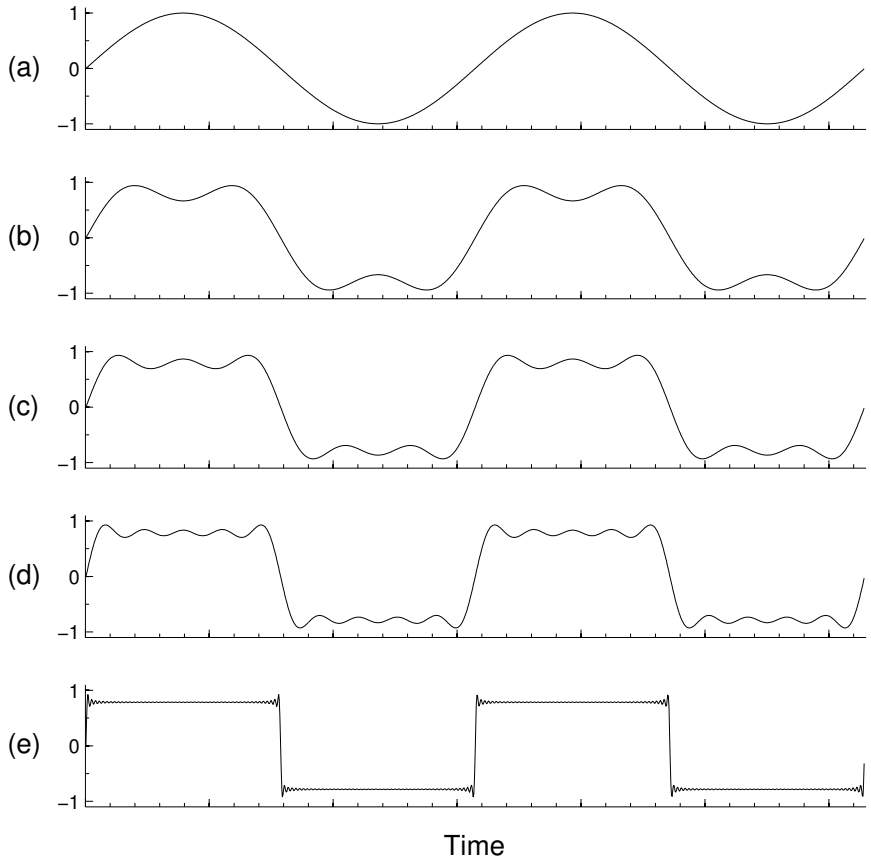


Figure 4.1: Stages of addition for the odd harmonic partials $(sin(x)+\frac{1}{3}sin(3x)+\frac{1}{5}sin(5x)+\ldots+\frac{1}{n}sin(nx))$ in the time-domain. (a) Fundamental waveform, (b) first and the third harmonics, (c) sum of odd harmonics through the fifth, (d) sum of odd harmonics through the ninth, (e) sum of odd harmonics up to the 101st creates a quasi-square wave.

The number of partials required to produce a complex sound can range from 20 to 50 (in this case we have used 30). Although the phase and the non-harmonic (stochastic) parts of a sound are not considered, an amplitude and a frequency envelope for each of the partials need to be controlled in order to absolutely define the evolution of a sound in time. It is clear that the high level of precision of this synthesis method results in a dramatic increase of the controllable parameters or, as Curtis Roads puts it, additive synthesis has 'a voracious appetite for control data' [Roads, 1996]. In other words, the dimensionality of the synthesis space is high. This generally constitutes a major drawback regarding the usability of additive synthesis. However, the fine control over specific partials was essential for this task and therefore additive synthesis

was considered ideal.

4.3 Algorithm

The two-section algorithm that was utilised for independent modification of the spectral centroid and the relative energy of the first three harmonics is described below.

4.3.1 Brightness modification with constant warmth

The modification of the spectral centroid position without affecting *warmth*[1] was achieved by altering the spectral distribution between the 4th and the 30th (last in our case) harmonics while preserving the overall energy in this region. For that purpose the above region was divided into two subgroups whose energy was altered according to the following procedure. The initial energies are given by Equation 4.2, 4.3, 4.4.

$$E_{27} = E_1 + E_2 \tag{4.2}$$

$$E_1 = \sum_{n=4}^{r} A_n^2 \tag{4.3}$$

$$E_2 = \sum_{n=r+1}^{30} A_n^2 \tag{4.4}$$

where E_{27} is the overall energy of the last 27 partials and r is the rounded harmonic 50% roll-off point[2] for the spectral region of the last 27 partials. Thus, the initial energies are close to equal ($E_1 \simeq E_2$).

Then the modification factors are calculated according to Equation 4.5, 4.6 and 4.7.

$$E_{27} = E_1 + E_2 = a^2 E_1 + b^2 E_2 \tag{4.5}$$

where a and b are the factors that multiply every harmonic amplitude in each subgroup. Based on Equation 4.5, b is expressed as a value of a (which is set by the experimenter) as shown in 4.6.

[1] Whenever the terms *brightness* and *warmth* are used in the text instead of their acoustic correlates they will refer to the expected auditory *brightness* and *warmth*.

[2] Mid-point of energy.

$$b = \sqrt{\frac{E_1 + E_2 - E_1 a^2}{E_2}} \tag{4.6}$$

The square root of Equation 4.6 introduces the following limitation for a.

$$a \leq \sqrt{\frac{E_1 + E_2}{E_1}} \tag{4.7}$$

It must also be stated that the above calculation does not require that E_1 equal E_2. However, the subgroups were divided based on the 50% roll-off point since in this way a more even modification of the spectral centroid around its initial value was achieved. Since both regions preserve their initial energies, this method does not alter the 'warmth ratio' while changing the position of the spectral centroid.

4.3.2 Warmth modification with constant brightness

The method used for *warmth* modification implemented a transformation of an existing signal using a modifying signal in the frequency domain. This transformation kept the spectral centroid constant while altering the relative energy of the first three partials. The modifier signal had the same spectral centroid as the original as shown in Equation 4.8.

$$SC_{org} = \frac{\sum_{n=1}^{N} nA_n^2}{\sum_{n=1}^{N} A_n^2} = SC_{mod} = \frac{\sum_{n=1}^{N} nX_n^2}{\sum_{n=1}^{N} X_n^2} \tag{4.8}$$

where A_n are the harmonic amplitudes of the original and X_n are the harmonic amplitudes of the modifier. Based on the following fraction identity:

$$\frac{a}{b} = \frac{c}{d} \Rightarrow \frac{a}{b} = \frac{a+c}{b+d} \tag{4.9}$$

we can construct a modified signal featuring the same spectral centroid as the original

$$\frac{\sum_{n=1}^{N} nB_n^2}{\sum_{n=1}^{N} B_n^2} = \frac{\sum_{n=1}^{N} n(A_n^2 \pm X_n^2)}{\sum_{n=1}^{N} (A_n^2 \pm X_n^2)} = SC_{org} = SC_{mod} \tag{4.10}$$

where $B_n = \sqrt{A_n^2 \pm X_n^2}$ are the harmonic amplitudes of the modified signal. Consequently, the above transformation provides a way of changing the spectral content of a signal without

altering its spectral centroid. For the purpose of this work, the modifier signal consists of three harmonics X_{n-1}, X_n and X_{n+1} (where n is the rounded normalised SC) that in essence create a formant around the normalised SC. X_{n-1} given X_n and X_{n+1} is calculated by Equation 4.12.

$$SC = \frac{(n-1)X_{n-1}^2 + nX_n^2 + (n+1)X_{n+1}^2}{X_{n-1}^2 + X_n^2 + X_{n+1}^2} \Longrightarrow \tag{4.11}$$

$$X_{n-1} = \sqrt{\frac{SC(X_{n+1} + X_n^2) - X_{n+1}^2(n+1) - X_n^2(n)}{n - 1 - SC}} \tag{4.12}$$

In this way a signal consisting of three harmonics and having a desired SC can be constructed. The effect of the algorithm on *warmth* is greater for signals having a normalised SC between 1.5 and 2.5 as in such a case it alters the first three partials of the sound.

4.4 Listening Test

Two identical pairwise dissimilarity rating listening tests were conducted in order to investigate the perceptual significance of the modifications applied by the algorithm. In addition, the tests examined the influence of the fundamental frequency on auditory *warmth* and *brightness*. The test was completed with a verbal elicitation part where selected pairs of stimuli had their differences verbally described.

4.4.1 Stimuli and Apparatus

The stimuli were generated by the application of the above algorithm to a parent timbre with an additive synthesis engine built in Max/MSP. Their spectrum was absolutely harmonic and consisted of 30 harmonics. The duration of all stimuli was chosen to be 1.6 seconds and the temporal envelope was the same for all samples (100 msec attack, 50 msec decay, 0.8 sustain level and 100 msec release) so that listeners could concentrate absolutely on spectral changes. Both rise and release were linear. The inter-stimulus interval (ITS) was 0.5 secs. Two groups of 12 stimuli were produced, differing only in fundamental frequency (220 Hz for the first and 440 Hz for the second group). The normalised SC of the parent timbre was selected to be 2.2 and was created using a *brightness* creation function [Jensen, 1999] shown in Equation 4.13.

For $A_n = B^{-n}$, where A_n is the amplitude of the *nth* harmonic, the normalised energy SC is calculated as follows:

$$SC_{norm} = \frac{\sum_{n=1}^{N\to\infty} n(B^{-n})^2}{\sum_{n=1}^{N\to\infty} (B^{-n})^2} \simeq \frac{B^2}{B^2-1} \tag{4.13}$$

and for a known SC_{norm}, B is calculated from Equation 4.14

$$B = \sqrt{\frac{SC_{norm}}{SC_{norm}-1}} \tag{4.14}$$

The reason for selecting the SC position in 2.2 was because it was desired for the *warmth* modification algorithm to affect only the amplitude of the first three harmonics and at the same time to obtain a reasonably *bright* sound. The positions of the twelve stimuli in the *warmth – brightness* feature space is shown in Figure 4.2.

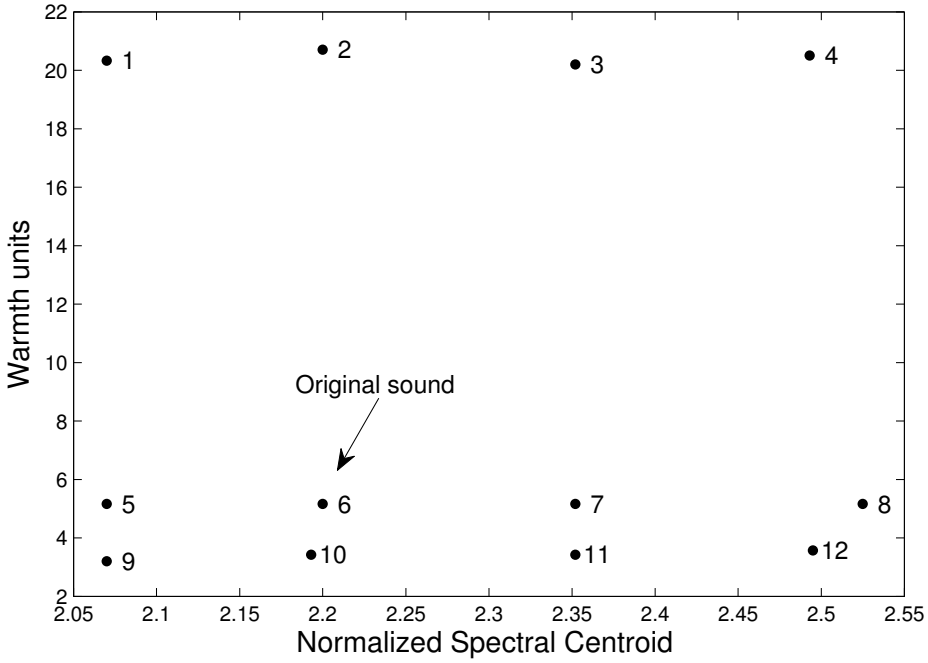


Figure 4.2: Feature space of the twelve stimuli.

The number of stimuli was set to 12 so that MDS analysis could produce up to a 3-D space according to the empirical rule of four stimuli per dimension [Green et al., 1989], while keeping the duration of the pairwise dissimilarity listening test relatively short. All stimuli were loudness equalised according to the experimenter's ear and only one out of 20 subjects reported difference in loudness between them. The stimuli were stored in PCM Wave format, at 16 Bit, 44.1Khz, in Mono.

The experiment was conducted through a Macbook Pro laptop with the AKG K 217 MK II

circumaural headphones, in a small acoustically isolated listening room. The interface of the experiment, part of which is presented in Figure 4.3, was built in Max/MSP.

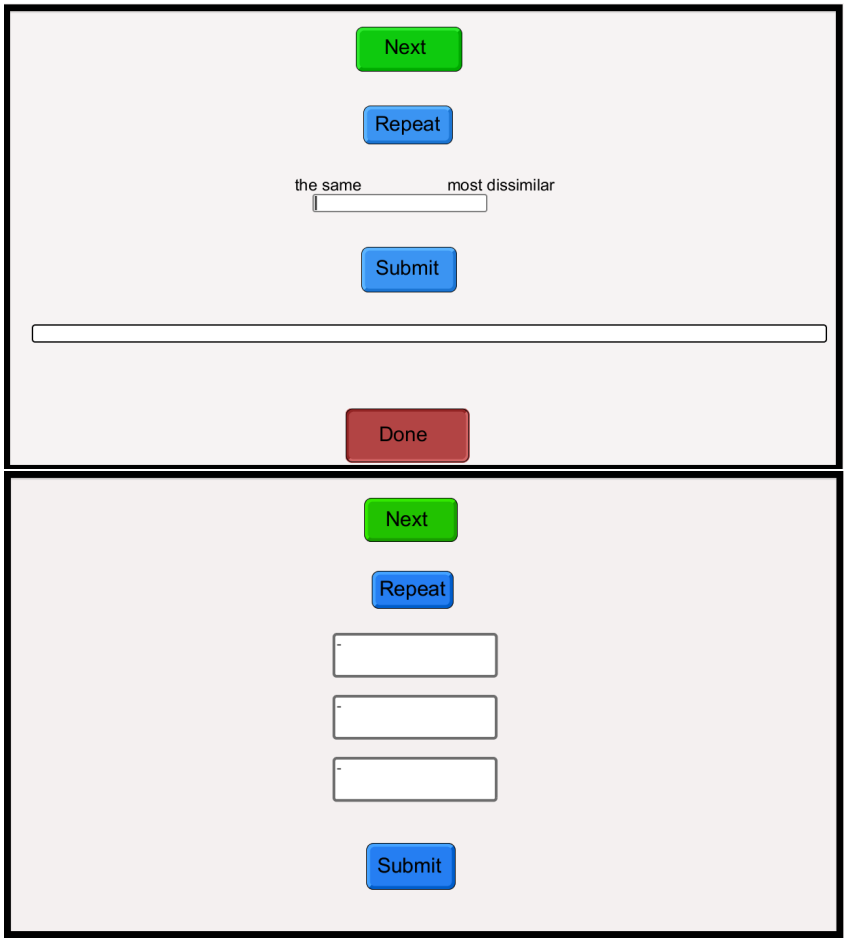


Figure 4.3: Sections of the listener interface. Pairwise dissimilarity test where listeners were asked to rate the dissimilarity between a pair of stimuli using the horizontal slider (Top). Verbal elicitation test where listeners were asked to insert up to three verbal descriptors for characterizing the difference between selected pairs of stimuli (Bottom).

4.4.2 Participants

Twenty participants (aged 23–40, 5 female) participated in the listening test. None of them reported any hearing loss and all of them had been practising music for 18 years on average (ranging from 8 to 30). Ten of them listened to the 220 Hz group of stimuli and the rest listened to the 440 Hz stimuli.

4.4.3 Procedure

Initially the listeners were presented with a familiarisation stage which consisted of random presentation of the stimuli in order for them to get a feel of the timbral range of the experiment. Subsequently, they performed a short training stage that consisted of five dissimilarity ratings. Finally, they undertook the complete pairwise dissimilarity test where they were randomly presented with all 78 combinations of pairs within the set. Comparisons of *same-sound* pairs were included as a measure of the validity of each listener. Listeners rated the differences of each pair using a hidden continuous scale with end–points marked as 'the same' and 'most dissimilar' as shown in Figure 4.3. They were also allowed to repeat the playback of each pair as many times as needed before submitting their rating.

4.4.4 Verbal Elicitation Test

The experiment was complemented with a verbal elicitation stage where listeners were presented with four selected pairs of stimuli in random order. The pairs used for this reason were the two diagonals of the quasi-rectangular feature space (1–12 and 4–9) as well as one pair on the 'warmth axis' (2–10) and one on the 'brightness axis' (5–8). The task of the listeners was to spontaneously insert up to three verbal terms that could describe how the second sound in the pair was different from the first (Figure 4.3). Again each pair could be played back as many times as necessary prior to submitting a description. The consistency of each listener's responses was tested by including two identical pairs of sounds in the test, thus increasing the overall number to six.

The overall listening test lasted approximately 35 minutes and participants were advised to take breaks if they felt signs of fatigue.

4.5 Results

4.5.1 MDS Analysis

The average dissimilarity matrices that were produced by the listener responses for both fundamental frequencies (F_0) were analysed through the MDS ALSCAL algorithm (see subsection 3.1.1) in SPSS[3]. The measures-of-fit of the MDS analysis were examined in order to determine

[3]All twenty subjects rated *same-sound* pairs as being identical (0 value) and as a result none was excluded from the analysis.

the optimal number of dimensions for this set of data. Table 4.1 shows the squared correlation index (RSQ) and the S-Stress tests (see subsection 3.1.1) for up to a 3-dimensional solution.

As the number of dimensions increases, RSQ will normally also increase while S-Stress will decrease. It is up to the researcher to decide the optimal dimensionality of the data based on the improvement of these measures versus increased complexity of the final solution. As shown in Table 4.1, the movement from 1-D to 2-D solution for the F_0 = 220 Hz case, results in an increase on the order of 0.1654 for the RSQ and also in a significant decrease of the S-Stress (0.1568). Adding a third dimension brings a negligible improvement to the measures (improvement < 0.05 for the RSQ). Thus, the optimal fit of the data appears to be a 2-D solution. The same can also be supported for the 440 Hz case, however with slightly worst results for the measures-of-fit values (Table 4.1).

Table 4.1: Measures-of-fit for the MDS solution of the 220 Hz and the 440 Hz pairwise dissimilarity tests. The scree plots (measure-of-fit value vs dimensionality) would have a 'knee' on the 2-D solution both for the RSQ and the S-Stress values which is a good indication that a 2-D space offers the optimal fit for this set of data.

	Dimensionality	**RSQ**	**RSQ improvement**	**S-Stress**	**S-Stress improvement**
220 Hz	1-D	0.78239	–	0.2842	–
	2-D	0.94784	0.16545	0.1274	0.1568
	3-D	0.95931	0.01147	0.09722	0.03018
440 Hz	1-D	0.81298	–	0.299	–
	2-D	0.88752	0.07454	0.1875	0.115
	3-D	0.91374	0.02622	0.1348	0.0527

The 2-D MDS spaces that were produced are shown in Figure 4.4. S_1-S_{12} represent the twelve stimuli (with S_6 being the original stimulus) and the arrows suggest an interpretation of the perceptual space. Indeed, S_1-S_4 change only in terms of the spectral centroid and S_1-S_5-S_9 change only in terms of the relative energy of the first three partials (see Figure 4.2). In the 220 Hz case the position of these two groups of stimuli resembles the feature space quite closely as they appear orthogonal in the perceptual space. Orthogonality is becoming weaker for stimuli with higher spectral centroids which are also perceived as being lower in the *warmth* dimension (S_8 and S_{12}). Additionally, for sounds with higher SC a decrease in *warmth* is also perceived as an increase in *brightness* (for example S_2-S_6 and S_3-S_7). This is an indication that for sounds with higher spectral content the modification of the SC and the warmth feature does not have a totally independent perceptual effect. Furthermore, the warmth feature relationship with

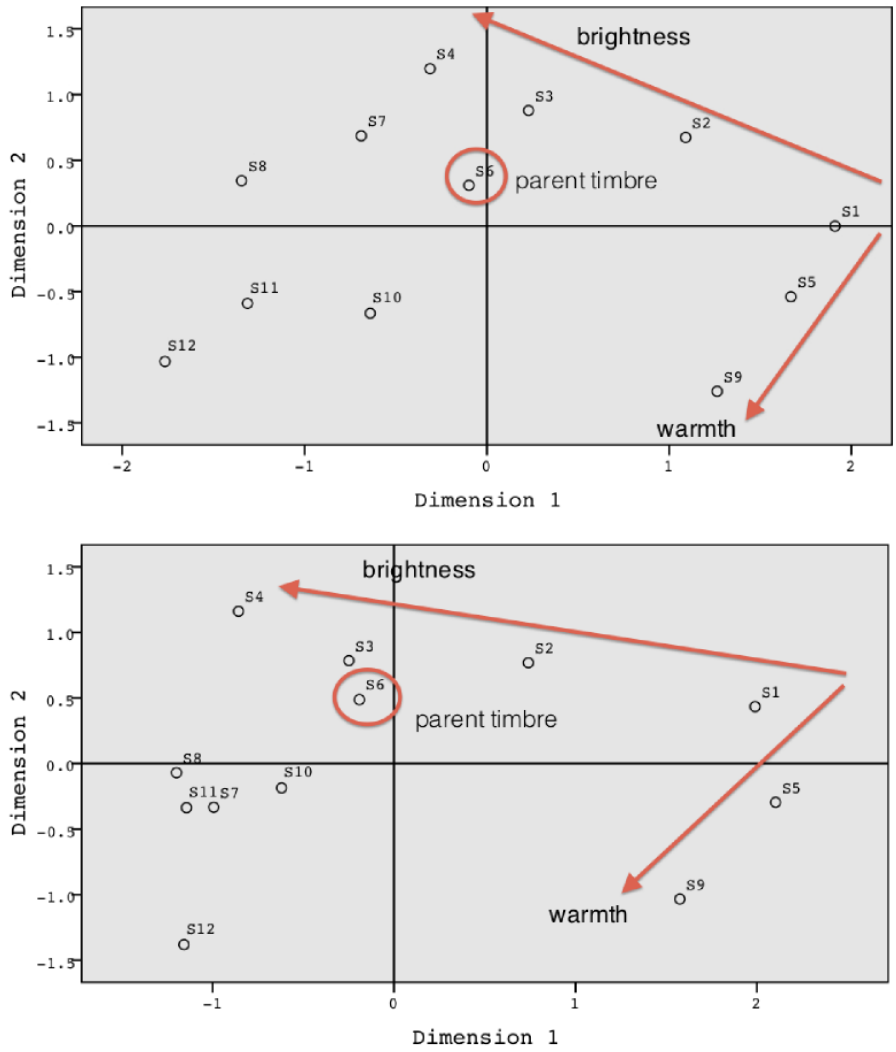


Figure 4.4: The two perceptual spaces created by the MDS analysis. The 220 Hz stimuli (Top) match the feature space better than the 440 Hz ones (Bottom). The brightness arrow shows the direction of SC increase and the warmth arrow shows the direction of *warmth* decrease.

perception seems to resemble a logarithmic one as the perceptual distances among S_1-S_5-S_9 are almost equal while in the feature space the S_1-S_5 distance is roughly 7.5 times larger than S_5-S_9. Finally, a widening of the perceptual space structure for sounds with higher spectral centroid is obvious as S_1-S_9 appear closer than S_4-S_{12} even though they are equidistant in the feature space.

For the 440 Hz case the matching of the feature space to the perceptual space is not that close. The S_1-S_5-S_9 group is again positioned somewhat independently from the the S_1-S_2-S_3-S_4 group but the angle between them is certainly less than 90°. Sounds with higher spectral centroid such as S_7-S_8-S_{10}-S_{11} are clustered together in the high *brightness*, medium *warmth* region. However, the space is still expanded for higher SCs (S_4-$S_{12} > S_1$-S_9).

Table 4.2: Pearson correlation coefficients between SC, warmth feature and Tristimulus 1, 2, 3 and the dimensions of the rotated MDS space for both F_0s. D_1 is parallel to the direction $S_1 \rightarrow S_5 \rightarrow S_9$ and D_2 parallel to the direction $S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow S_4$. (*: p<0.05, **: p<0.01), ***: p<0.001)

	Dimensions	**SC**	**Warmth**	**T1**	**T2**	**T3**
220 Hz	D1	-0.28	**0.80****	-0.59*	**0.935*****	**-0.81****
	D2	**0.91*****	-0.176	-0.44	-0.58*	**0.83*****
440 Hz	D1	-0.30	**0.82****	-0.57	**0.90*****	**-0.77****
	D2	**0.87*****	-0.49	-0.53	-0.38	**0.79****

The Pearson correlation coefficients between the perceptual space dimensions and some spectral features extracted from the sounds are shown in Table 4.2. T_1, T_2 and T_3 stand for Tristimulus 1, 2 and 3 [Pollard and Jansson, 1982] which are shown in Equations 3.47, 3.48 and 3.49.

The two MDS spaces were rotated clockwise by 60° and 72° in order to achieve an alignment between the *warmth* and *brightness* axes with dimensions 1 and 2 correspondingly. It is clear that D_2 is highly correlated with the spectral centroid. D_1 on the other hand seems to have a significant correlation with the warmth feature but is even stronger correlated with T_2. T_3 has both a positive correlation with D_2 and a negative correlation with D_1.

4.5.2 Verbal Elicitation

Tables 4.3 and 4.4 show the results of the verbal elicitation part of the test for the two different fundamental frequencies.

All twenty subjects were consistent with their verbal judgements between identical pairs. They usually did not use the exact same verbal descriptors for both cases but the context was always the same. The groupings were made based on semantic relevance and according to the groupings in Williams and Brookes [2010], Štěpánek [2006], Howard et al. [2007].

The pair S_1-S_{12} represents a difference from the maximum *warmth* and minimum *brightness* to the maximum *brightness* and minimum *warmth*. The most prominent group of answers is the one that includes the descriptor 'bright'. This is clearer for F_0 = 440 Hz and indicates that the simultaneous increase of SC and decrease of the warmth feature results in an increase of auditory *brightness*. Only one out of forty one answers was 'less warm' even though the warmth feature had roughly decreased by 80% of its initial value and SC had only increased by 25%.

The S_4-S_9 pair provides even more revealing results. The movement for this pair is from

Table 4.3: Verbal elicitation results for the pairs of the F_0 = 220 Hz group. Words in bold indicate the word with higher frequency of appearance within the group.

	Groups of adjectives used to describe differences between sounds	**Number of occurrences**	**Percentage of the total number of answers**
how S_1 differs from S_{12}	**bright**, clear, trebly	8	40%
	fuzzy, crackly, buzzy, harsh, robotic, less round	6	30%
	small, thin, tight	3	15%
	various	3	15%
how S_4 differs from S_9	**dull**, gloomy, damp, muted, closed	7	28%
	soft, smooth	6	24%
	warm, round	5	20%
	full, dense	3	12%
	various	4	16%
how S_2 differs from S_{10}	**bright**, treble	9	56%
	big, full, open	3	19%
	harsh, buzzy	2	12.5%
	warm, less pleasant	2	12.5%
how S_5 differs from S_8	**bright**, nasal, clear, treble	11	55%
	thin	3	15%
	various	6	30%

maximum *brightness* and *warmth* to minimum *brightness* and *warmth*. The results for both F_0s do not suggest a unique prominent group but rather three groups of descriptors that have the highest frequency of appearance. Sound S_9 is generally rated as being warmer, duller or darker and softer or smoother. This fact implies that the perception of *brightness* overshadows the perception of *warmth*, and that *warmth* might be the perceptual antonym of *brightness*. Indeed, no one rated S_9 as being less *warm*. On the contrary, many participants actually described it as being *warmer*. This shows a discrepancy between the suggested warmth feature and the actual perception of *warmth*, and also a high level of overlap between *brightness* and *warmth*.

The S_2-S_{10} pair represents a movement from maximum to minimum *warmth* having a constant SC position. Tables 4.3 and 4.4 show that the *brightness* group predominates with very similar results for both F_0s. This result reveals that participants rated S_{10} as *brighter*, even though the position of the SC was exactly the same with S_2. This agrees with the MDS spaces that position S_{10} away from S_2 both in *warmth* and *brightness* direction. Despite the fact that the distance in *warmth* is greater than the distance in *brightness*, it is the latter that is spontaneously verbalised. The fact that no one responded 'less warm' or 'colder' needs to be highlighted and contributes to the hypothesis of *warmth* being a perceptual antonym for *brightness*.

Table 4.4: Verbal elicitation results for the pairs of the F_0 = 440 Hz group. Words in bold indicate the word with higher frequency of appearance within the group.

	Groups of adjectives used to describe differences between sounds	**Number of occurrences**	**Percentage of the total number of answers**
how S_1 differs from S_{12}	**bright**, **sharp**, less muffled, nasal, edgy	13	65%
	ring, harsh, metallic	3	15%
	thin, reedy, less brass	3	15%
	less warm	1	5%
	full	1	5%
how S_4 differs from S_9	**warm**, round	7	30%
	dark, **dull**, less nasal	7	30%
	muffled, smooth, less harsh	5	22%
	thick, more body	2	9%
	various	2	9%
how S_2 differs from S_{10}	**bright**, less dull, nasal	10	53%
	rich, full, more harmonics	3	17%
	thin	2	10%
	harsh, punchy	2	10%
	various	2	10%
how S_5 differs from S_8	**bright**, nasal, penetrating, sharp, edgy	11	55%
	harsh	2	10%
	less round, less warm	2	10%
	various	5	25%

Finally, the S_5-S_8 pair represents an increase of SC while keeping the warmth feature constant. The results are again quite similar for both F_0s and indicate that the *brightness* group is the most prominent but at the same time there is a significant number of responses that are not grouped. This is an expected result that confirms previous research regarding the perceptual relevance of spectral centroid [e.g. Lichte, 1941, von Bismarck, 1974a, McAdams et al., 1995, Schubert et al., 2004, Williams and Brookes, 2007].

4.6 Discussion

An algorithm for the independent modification of the spectral centroid and the relative energy of the first three partials of a harmonic sound was designed and implemented through a Max/MSP additive synthesis engine. The perceptual validity of these two features together with the potential influence of the fundamental frequency were investigated through two pairwise dissimilarity listening tests.

The 2-D spaces that were produced by MDS analysis demonstrate a relatively good matching

between the feature space and the perceptual space for F_0 = 220 Hz. It is also evident that for low spectral centroids the modification of these two features is perceived independently and that for higher spectral centroids there seems to be a degree of overlap between them. For F_0 = 440 Hz, the matching between the two spaces worsens significantly but there still is evidence of perceptual independence for lower SCs. The correlation between the rotated axes of the space (so that they coincide with what seems to be the basic directions of movement on the MDS space) and some spectral features were calculated. A strong correlation between D_2 and SC and a correlation between the warmth feature and D_1 were revealed for both fundamentals. The difference between the MDS spaces, suggests that fundamental frequency might have some influence on the perception of these particular modifications. Further research is mandated towards this direction. At this point, it must be stated that Tristimulus 2 features the strongest correlation with D_1 and also that Tristimulus 3 features a quite strong negative correlation with D_1. This is a sign that T_2 or/and T_3 might influence the listeners' judgements more than the warmth feature.

Although the MDS analysis showed that a degree of perceptual independence among sounds with different *warmth* and SC does exist, the verbal elicitation experiment did not support semantic independence. 'Bright' was the most prominent semantic descriptor that was elicited through the free response test for describing an increase in the SC, decrease in *warmth* and a combination of the two. For the decrease in *warmth* and SC the terms varied and the three most prominent terms were dull, warm and soft.

The results of this work question the claim that the relative energy of the first three partials is an adequate acoustic correlate for auditory *warmth*. Furthermore, they seem to agree with previous findings that supported the existence of a degree of overlap between auditory *brightness* and *warmth* [Howard et al., 2007, Ethington and Punch, 1994, Pratt and Doak, 1976]. Another interesting finding is the fact that sounds with the same spectral centroid are rated as differing in *brightness*. This implies that auditory *brightness* is not determined merely by the spectral centroid position.

4.7 Conclusion

This chapter focused on a very specific problem of musical timbre semantics and its acoustic correlates. The findings did not confirm the previously suggested acoustic correlate for auditory *warmth*. Furthermore, the widely accepted notion that spectral centroid is the acoustic correlate

for auditory *brightness* was also questioned since sounds with the same spectral centroid had been rated as differing in *brightness*.

Auditory semantic attributes are most likely multifactorial. Thus the physical correlates of a semantic dimension cannot be revealed merely by examining separate audio descriptors. The presented findings have failed to identify an undoubted acoustic correlate of any semantic dimension, have failed to inform us of the exact relationship between auditory *brightness* and *warmth* (i.e., are they synonyms, antonyms or completely independent?) and have even failed to demonstrate the salience of *warmth* as a timbral semantic descriptor.

As mentioned in the introduction, one of the main objectives of this work is the development of a common semantic framework for musical timbre description. All the above have demonstrated that a more holistic approach is required to pursue this goal. It was made clear that we should first aim at identifying the most significant semantic dimensions of timbre and subsequently associate them with physical properties of the sound signal. To this end, we designed a new experiment of high ecological validity[4] which examined a number of commonly used musical sounds instead of specifically synthesised samples. Following the experience we gained from the current experiment, but also from informal discussions with composers and professional musicians, we came to the conclusion that even though the use of adjectives for timbre description is intuitive in general, spontaneous verbalisation might be problematic. We have therefore decided that the best approach would be to provide our participants with a large predefined vocabulary of semantic descriptors enhanced with the option of additional free verbalisation.

The following two chapters, which constitute the core of this thesis, investigate the semantics of musical timbre between two different language populations (Greek and English) and examine the relationship of the identified semantic space with perception.

[4] A research study is ecologically valid when it constitutes a good approximation of the real world.

Chapter 5

Semantic dimensions of musical timbre: investigating language dependence and their acoustic correlates

---

5.1 Introduction

This chapter will present an experiment designed to investigate the influence of language on semantic descriptions of timbre. In two separate listening tests, native Greek and English speaking participants were asked to describe 23 musical instrument tones using a predefined vocabulary of 30 adjectives. This allowed for direct comparison between the two different linguistic groups. A combination of continuant and impulsive stimuli of both acoustic and synthetic nature that also varied in pitch has been rated through Verbal Attribute Magnitude Estimation (VAME) in order to reach generalizable conclusions regarding timbre semantics.

A data reduction methodology combining Cluster Analysis (CA) and Factor Analysis (FA) (see subsections 3.1.2 and 3.1.3 respectively) was followed in order to identify the salient semantic dimensions of timbre for both languages. FA assumes linear relationships between the variables under study. This, however, is not always guaranteed to be the case when analysing semantic variables. Transformations (rank ordering, optimal spline ordinal) applied to the data have helped to investigate potential nonlinear relationships between the examined verbal attributes. It was demonstrated that the proper treatment of such nonlinearities can enhance the robustness of the resulting semantic space.

Finally, the acoustic correlates of the major semantic dimensions were identified through a correlation analysis. This identification has been a matter of ambiguity between various studies

[e.g. von Bismarck, 1974b, Ethington and Punch, 1994, Faure et al., 1996, Disley et al., 2006]. The association of timbre semantics with certain physical characteristics of sound is highly desirable as it contributes towards a better understanding of timbre perception and facilitates the development of intuitive sound processing applications.

5.2 Method

A listening test based on a modification of the verbal attribute magnitude estimation (VAME) method was designed and conducted. VAME was preferred for the purpose of this study because, unlike the semantic differential, it reduces potential biases associated with assumptions concerning synonym and antonym relationships between the verbal labels for the rating scales. As a trade off, VAME requires double the number of verbal variables for the same number of adjectives in comparison to the semantic differential (see section 2.5).

The listeners were provided with a preselected list of 30 verbal descriptors[1] (Fig. 5.1) in their native language and were asked to describe the timbral attributes of 23 sound stimuli by choosing the adjectives they believed were most salient for each stimulus. No limit was imposed on the number of adjectives that could be used by each participant for each description. The verbal descriptors provided were intended for the description of sound impressions [Wake and Asahi, 1998] and were selected among adjectives that are commonly found in musical timbre perception literature [Ethington and Punch, 1994, von Bismarck, 1974a,b, Faure et al., 1996, Disley et al., 2006]. The collection of terms is given in Table 5.1. Once a listener chose a descriptor he or she was asked to estimate the value that corresponded to the sound on a scale anchored by the full extent of the verbal attribute and its negation, such as 'not sharp-very sharp' (Fig. 5.1). This rating was input using a horizontal slider with a hidden continuous scale ranging from 0 to 100. A source of criticism regarding the provision of a predefined vocabulary is that the set of verbal attributes does not always correspond to descriptors that the participants would chose spontaneously [Donnadieu, 2007]. To alleviate such issues the listeners were allowed to freely propose up to three additional adjectives of their own choice to describe each stimulus.

[1]The selection of the adjectives was based on the existing literature on timbre semantics as described in section 2.5.

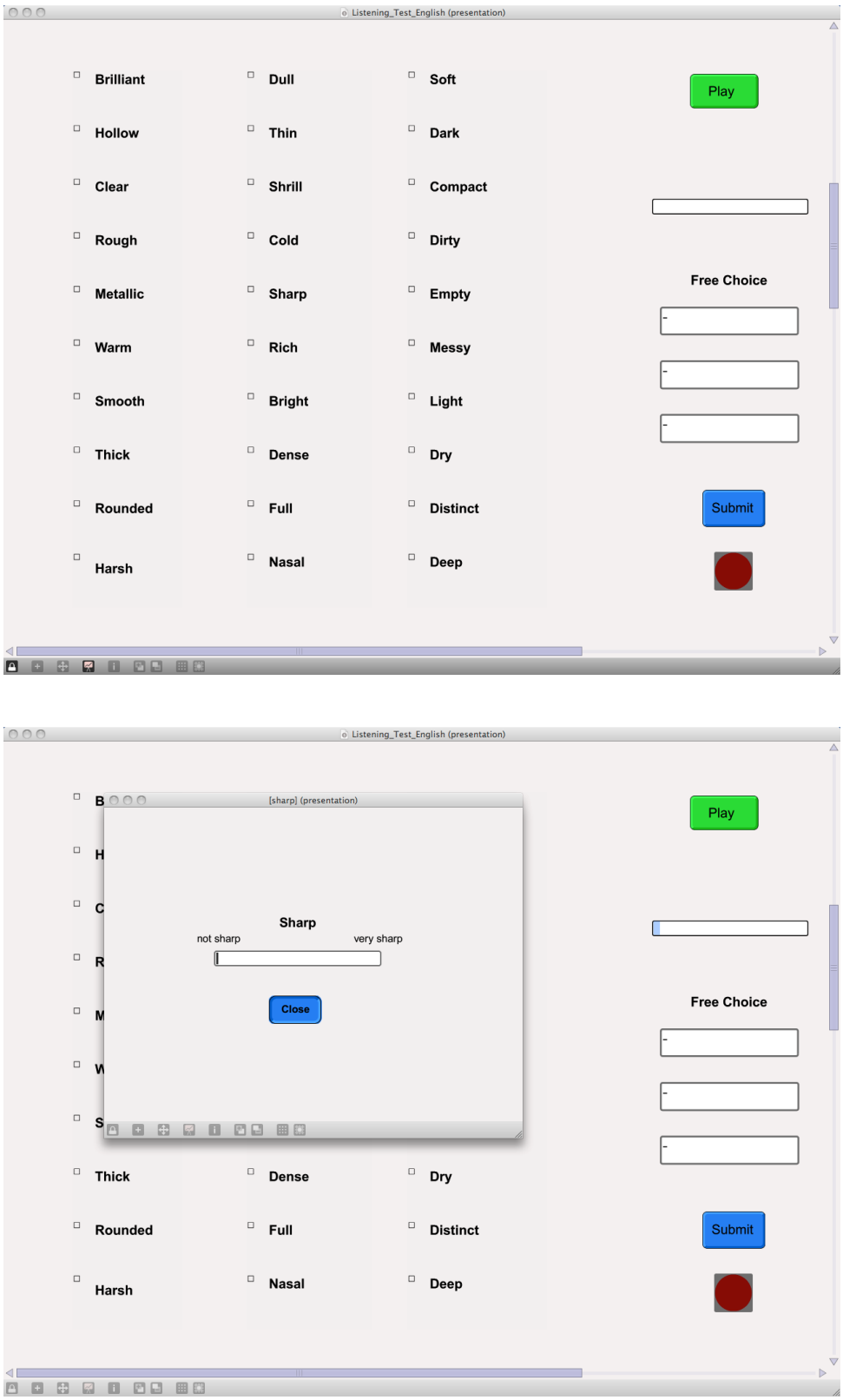


Figure 5.1: The Max/MSP customised interface of the subjective evaluation listening test (top) and the pop up window that appeared each time the participant picked up an adjective (bottom).

Table 5.1: Spearman correlation coefficients between the 30 equivalent semantic variables (descriptors) of the two languages (italics: $p < 0.05$, bold: $p < 0.01$). The Greek equivalent terms as translated by a linguist appear in parentheses.

Descriptor	Correlation	Descriptor	Correlation
Brilliant (Λαμπερός)	**0.769**	Sharp (Οξύς)	**0.665**
Hollow (Υπόκωφος)	-0.077	Rich (Πλούσιος)	0.372
Clear (Καθαρός)	**0.543**	Bright (Φωτεινός)	**0.802**
Rough (Τραχύς)	**0.819**	Dense (Πυκνός)	**0.803**
Metallic (Μεταλλικός)	**0.813**	Full (Γεμάτος)	**0.698**
Warm (Ζεστός)	**0.732**	Nasal (Ένρινος)	**0.730**
Smooth (Μαλακός)	**0.847**	Soft (Απαλός)	**0.620**
Thick (Παχύς)	**0.801**	Dark (Σκοτεινός)	**0.599**
Rounded (Στρογγυλεμένος)	**0.860**	Compact (Συμπαγής)	0.018
Harsh (Σκληρός)	**0.819**	Dirty (Βρώμικος)	**0.773**
Dull (Θαμπός)	0.399	Empty (Άδειος)	0.020
Thin (Λεπτός)	**0.779**	Messy (Τσαλακωμένος)	*0.521*
Shrill (Διαπεραστικός)	**0.853**	Light (Ελαφρύς)	**0.668**
Cold (Ψυχρός)	*0.506*	Dry (Ξερός)	**0.610**
Distinct (Ευδιάκριτος)	*0.520*	Deep (Βαθύς)	**0.854**

5.2.1 Stimuli and Apparatus

Aiming to promote ecological validity, a set of 23 sounds drawn from commonly used acoustic instruments, electric instruments and synthesisers and with fundamental frequencies varying across three octaves was selected. The following 14 instrument tones come from the McGill University Master Samples library [Opolko and Wapnick, 2006]: *violin, sitar, trumpet, clarinet, piano* each at A3 (220 Hz), *Les Paul Gibson guitar, baritone saxophone B flat* each at A2 (110 Hz), *double bass pizzicato* at A1 (55 Hz), *oboe* at A4 (440 Hz), *Gibson guitar, pipe organ, marimba, harpsichord* each at G3 (196 Hz) and *French horn* at A#3 (233 Hz). A *flute* recording at A4 was also used along with a set of 8 synthesiser and electromechanical instrument sounds: *Acid, Hammond, Moog, Rhodes piano* each at A2, *electric piano (rhodes), Wurlitzer, Farfisa* each at A3 and *Bowedpad* at A4.

Musical timbre studies usually restrict the sound stimuli to a fixed fundamental frequency (F_0). The reason why we have chosen to relax this restriction was to stimulate a wider range of verbal descriptions, to enhance generalisation of the findings and to also investigate the influence of F_0 on the semantic dimensions of musical timbre. Marozeau et al. [2003] and Marozeau and de Cheveignè [2007] have investigated this influence as well. Furthermore, Alluri and Toiviainen [2010] and Alluri et al. [2012] have shown that listeners can consistently rate short musical ex-

cerpts of varying key and rhythm on semantic scales. Since the task of this experiment was the assignment of a value of a semantic descriptor rather than a strictly controlled pairwise comparison, the stimuli were not required to be of equal duration either. Durations ranged from 3 to 8 seconds depending on the nature of the instrument (continuant or impulsive). Nevertheless, sound samples were equalised in loudness in an informal listening test within the research team. The RMS playback level was set between 65 and 75 dB SPL (A-weighted). Eighty three percent (83%) of the Greek participants found that level comfortable for all stimuli and 78% reported that loudness was perceived as being constant across stimuli. For the English participants these values were 93% and 85%, respectively.

The listening test was conducted in acoustically isolated listening rooms. Sound stimuli were presented through the use of a laptop computer with an M-Audio (Fast Track Pro USB) external audio interface and a pair of Sennheiser HD60 ovation circumaural headphones.

5.2.2 Participants

A first linguistic group consisting of 41 native Greek speakers (aged 19-55, mean age 23.3, 13 male) and a second one consisting of 41 native UK English speakers (aged 17-61, mean age 29.6, 28 male) participated in the listening test. None of the listeners reported any hearing loss and they had been practicing music for 13.5 (Greek) and 18.8 (English) years on average, ranging from 5 to 35 (Greek) and from 4 to 45 (English). There was also a prerequisite that participants did not have sound related synaesthesia or absolute pitch, as such a condition could affect the results due to pitch variation within the stimulus set. Participants were students of the Department of Music Studies of the Aristotle University of Thessaloniki, researchers from the Centre for Digital Music at Queen Mary University of London, students of the Royal College of Music and of the Music Department of Middlesex University in London.

5.2.3 Procedure

Listeners became familiar with the timbral range of the experiment during an initial presentation of the stimulus set (random order). On each trial of the experimental phase, participants were presented with one sound stimulus. They could listen to it as many times as required before submitting their ratings. The sounds were presented in random order and listeners were advised to use as many of the provided terms as they felt were necessary for an accurate description of each different timbre, and also to take a break when they felt signs of fatigue. The overall

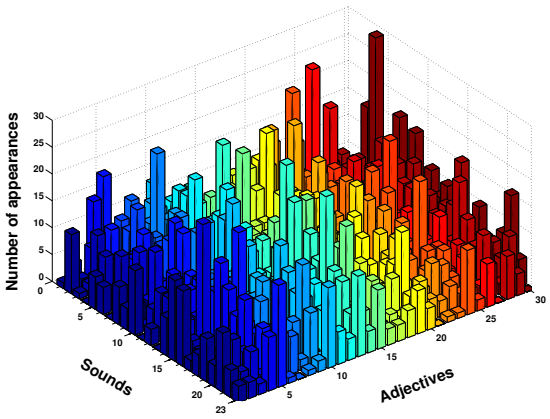
listening test procedure, including instructions and breaks, lasted approximately 45 minutes.

5.2.4 Cluster Analysis, Factor Analysis and CATPCA transformation

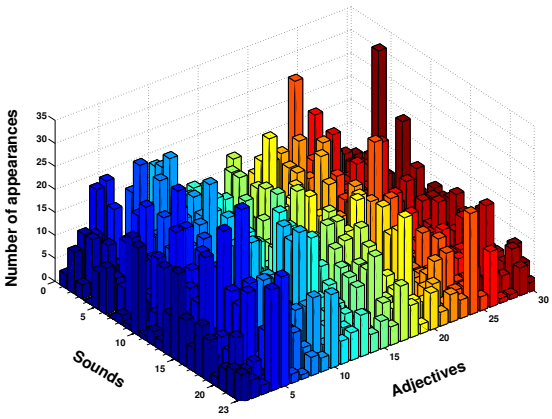
Two statistical analysis techniques were applied to the data in order to reach conclusions regarding the salient semantic dimensions of timbre. Cluster Analysis [Romesburg, 2004] indicated groups of semantically related verbal descriptors while Factor Analysis (FA) [Harman, 1976] uncovered the latent structure of the inter-correlated semantic variables.

As already mentioned, an important element of the analysis in this work is the fact that it allowed for the possible existence of nonlinear relationships between the measured verbal attributes. That is, within the framework of FA, the constraint on strict linear relations between variables was relaxed by anticipating necessary optimal transformations of the original variables along with data reduction. For this reason, an optimal transformation of the variables through CATegorical PCA (a readily available technique as a computational realisation within the SPSS 20 suite.) was employed. More details about cluster and factor analyses and CATPCA transformation are provided in the methods section 3.1.

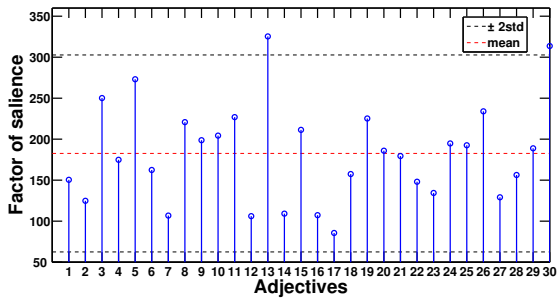
However, CATPCA deals with a PCA problem, which is maximising described variance in contrast to FA which seeks for identification of underlying structure. Also, the software implementation in SPSS does not offer additional rotation techniques for the derived solution. Since our goal was to address the problem of timbre description under a factor analytic approach rather than with PCA, we followed a hybrid approach. We first employed CATPCA in order to obtain nonlinear transformations of the original variables and we then used the transformed variables to conduct FA with rotation. Our approach aims at capturing and linearising possible nonlinear relationships among original variables, thus improving the performance of the derived FA solution, while at the same time leaving the solution intact when linear relationships prevail. The usefulness of the approach is tested both by inspection of the form and extent of possible nonlinear transformations (see Appendix A) of the original variables, and by the impact of such a 'linearisation' upon the final solution (compared to the typical 'untransformed' FA).



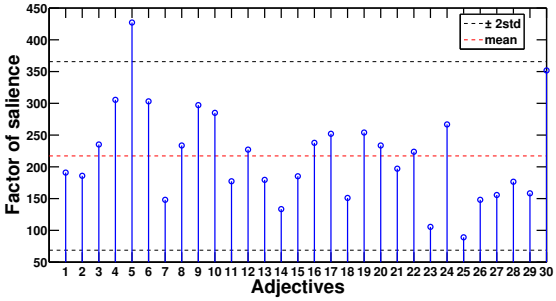
(a) Greek



(b) English



(c) Greek



(d) English

Figure 5.2: Number of appearances for each adjective per sound for Greek (a) and English (b) listeners. Factor of salience for the Greek (c) and English (d) adjectives. No adjective had a factor of salience less than twice the standard deviation from the mean and therefore all adjectives were considered salient.

5.3 Analysis and Results

5.3.1 Measure of salience for each adjective

The 3D bar charts presented in Figure 5.2 (a) and (b) show the number of appearances of each adjective per sound for both linguistic groups. Prior to applying statistical analysis techniques to the data of the two groups, the salience of the descriptive adjectives was tested using the following criterion that is based on the number of times that each adjective was selected by the participants:

$$S(i) = \sum_{n=1}^{23} a_n(i) + \frac{max(\overline{a(i)})^k}{\sum_{n=1}^{23} a_n(i)} \tag{5.1}$$

where $S(i)$ is the factor of salience for each adjective i, $a_n(i)$ is the number of times a certain adjective i has been chosen by all the participants for describing a particular sound sample n and $\overline{a(i)}$ is the $(1,23)$ vector that contains the number of appearances corresponding to adjective i for the 23 sounds. This factor takes into account a combination of both the overall number of appearances and the maximum number of these appearances for each adjective. This is because even if an adjective has only a small number of overall appearances among all sound samples, a single high maximum at one particular sound can suggest that this adjective is still meaningful. Therefore, a balance between the two terms of Equation 5.1 needed to be maintained. As a result, the power k, to which the maximum number of appearances is raised, was heuristically set to 3 after observation of the metric attitude in relation to the number of sounds. The calculation of S for all the adjectives revealed that S was always greater than or equal to the mean minus two standard deviations for both groups of listeners. Therefore, no adjective could be characterised as a non-significant outlier and none was discarded at that stage.

The magnitude ratings for each verbal descriptor and each musical timbre were averaged across the 41 participants in each of the language groups. Thirty seven percent (37%) of the Greek participants inserted at least one extra verbal descriptor, thus providing 31 additional terms. However, only 8 of these terms where mentioned more than once, and only 6 were mentioned by more than one participant. Sixty six percent (66%) of English participants used at least one extra term, thus providing 117 additional verbal descriptors. Thirty three of these terms were inserted more than once, and 27 were used by more than one participant. The extra terms are presented in Table 5.7 and discussed in subsection: *Inter-linguistic relationships* (5.3.4).

5.3.2 Statistical analysis

Step-by-step data reduction methodology

As previously explained, factor analysis was the favoured data reduction method applied to identify the salient semantic dimensions of musical timbre for both English and Greek. In FA, a mild multicollinearity between variables (in this case verbal descriptors) is generally desirable and for this reason variables that either correlate very highly (extreme multicollinearity) or variables that are not correlated with the rest of the group are discarded prior to the analysis. The steps followed towards data reduction within each linguistic group are summarised below:

- A hierarchical cluster analysis (centroid linkage) based on squared Euclidean distances over the verbal descriptors (see subsection 3.1.2), identified the major clusters and outliers among them. The outliers were adjectives that could not be grouped with other adjectives as they appeared to have many instances of low inter-correlation coefficients. As a consequence such variables were discarded based on an observation of the dendrogram. For example, the terms *empty* in Greek and *cold* and *compact* in English all form a cluster on their own in dendrograms 5.4 c) and 5.4 d) respectively and were thus removed.

- In order to further reduce the number of verbal descriptors a preliminary Factor Analysis was performed within each cluster and a non-orthogonal oblique rotation[2] of the extracted factors was employed. The criterion used for deciding the number of factors (eigenvalues ≥ 1) resulted in either two or three factor solutions in all cases. The adjectives with extracted communalities < 0.6 were then discarded as the communality measures the percentage of variance in a given variable explained by all the factors jointly. This criterion ensured that only the verbal descriptors that were adequately explained by the model for each cluster were retained.

- The correlation matrix of the remaining adjectives was inspected and extremely multicollinear ($r > 0.8$) verbal descriptors were removed.

- The descriptors selected in the preliminary stage were then subjected to a FA, again applying oblique rotation to increase the interpretability of the factors. The descriptors featuring

[2]When a solution features two or more factors, the possible orientations are infinite. Factor rotation provides the solution with the best simple structure. This is achieved by maximising the already large factor loadings and minimising the small ones. Non-orthogonal rotation does not presuppose that factors are uncorrelated, thus provides more accurate and realistic solutions.

communalities < 0.6 were again discarded and the remaining set of descriptors was subjected to a final FA. The final data reduction step uses *factor loadings* as a criterion for labelling the major factors.

Nonlinear transformation of the variables

A non-metric factor analytic approach has been shown to relax the strict assumption of linear relationships between variables allowing for the investigation of monotonic nonlinearities [Woodward and Overall, 1976]. Following this approach, a preliminary analysis of the English group data, published in Zacharakis et al. [2012], showed that a simple rank ordering transformation of the verbal variables explained a larger amount of variance with fewer dimensions compared to the untransformed case. Table 5.2 shows the percentages of factorial and total variance explained by the data reduction methodology described above for original and rank transformed data. It is evident that there is both a small increase (3%) of the total explained variance and a significantly higher concentration of the accounted variance (additional 7.6%) in the first two factors for the transformed variables. It was assumed that this was an indication that existing nonlinearities among the perceptual variables have been more efficiently modelled by the non-metric approach. Based on this finding, an optimal spline [3] ordinal transformation (2nd degree and 2 interior knots) performed by the CATPCA module of SPSS suite has been applied to the variables. The number of categories was set to 7. This transformation has additionally contributed towards addressing issues with strongly skewed data. Figure 5.3 shows two indicative nonlinear transformation plots obtained by the CATPCA optimization as an example of the shape of transformations applied to the variables. All 60 transformation plots (30 adjectives by 2 languages) are presented in Appendix A.

Table 5.2: Total and factorial variance explained prior the non-orthogonal rotation for the original and rank transformed variables.

	1st Factor	**2nd Factor**	**3rd factor**	**Total**
Original	42.8%	24.5%	9.8%	77.1%
Rank Transf.	46.4%	28.5%	5.2%	80.1%

Figures 5.4a and 5.4b show the dendrograms of the original adjectives and Figures 5.4c and 5.4d show the dendrograms of the transformed adjectives as resulting from the application of

[3] A spline is a piecewise polynomial function defined by a degree or order (degree plus one) and a set of interior knots where the polynomial pieces connect.

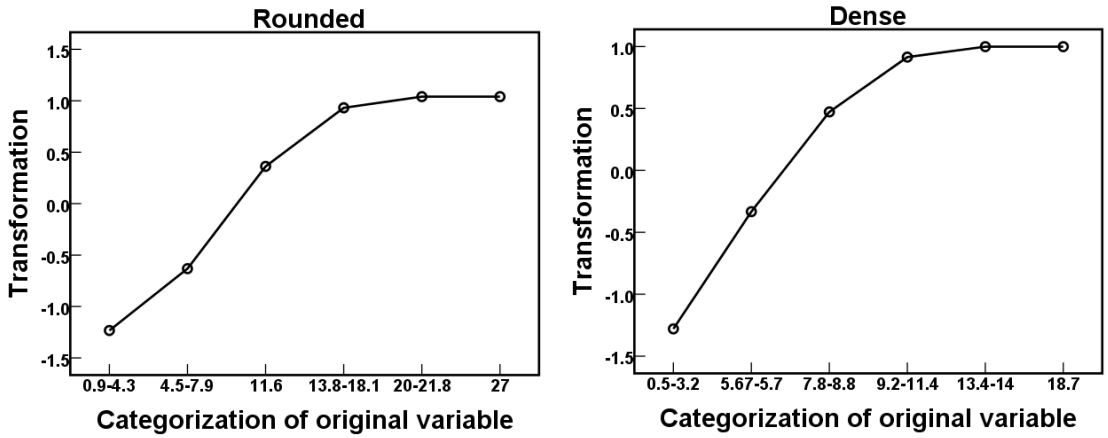
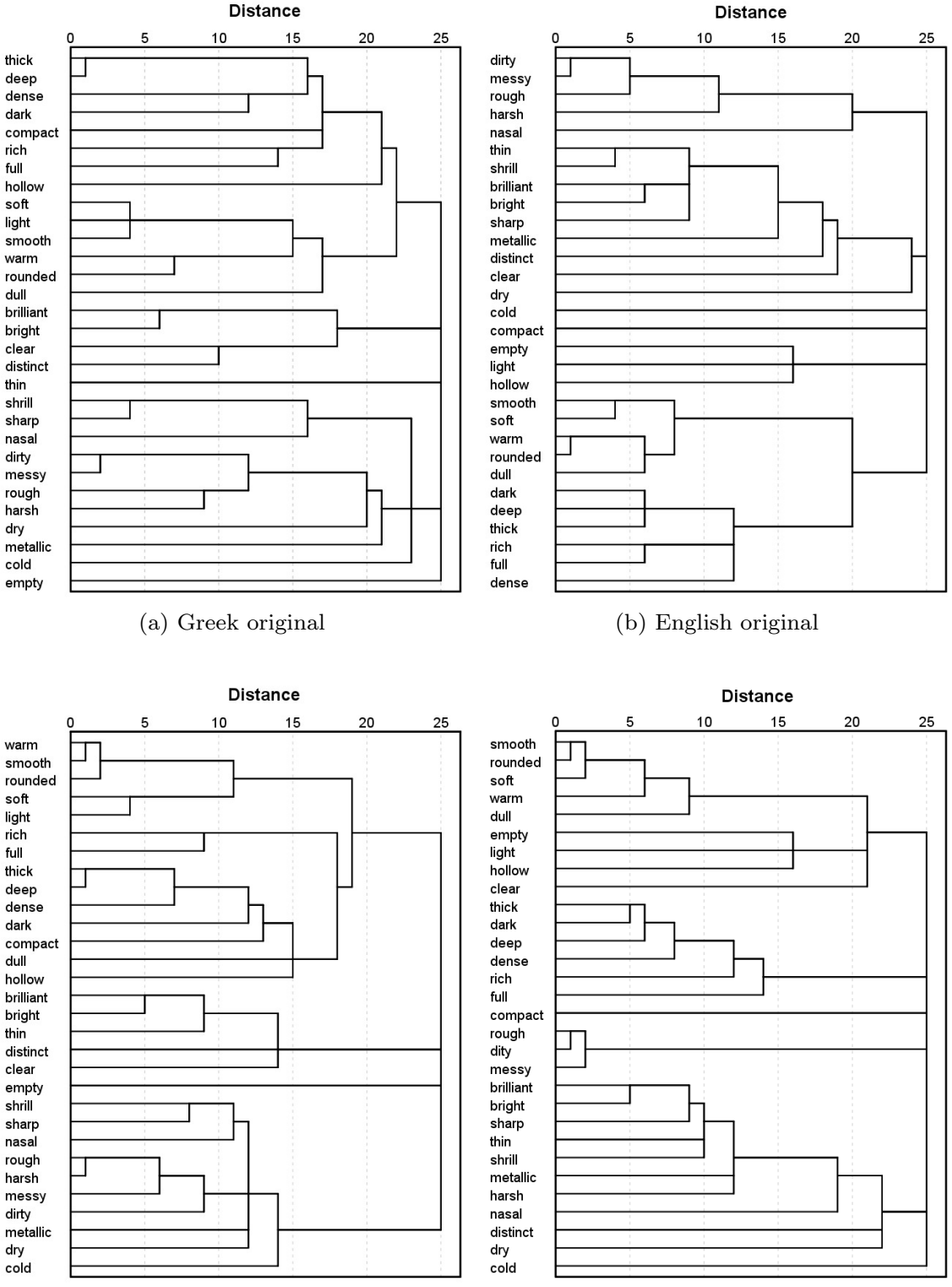


Figure 5.3: Indicative optimal nonlinear transformations of original variables. Rounded (Greek) on the left and Dense (English) on the right. The abscissa represents the categories in which the variable is separated (in this case six) and the ordinate represents the value that is assigned to each category by the algorithm.

cluster analysis to both linguistic groups. In the original dendrograms, the absence of clearly defined clusters reflects the lack of cohesive groups among the adjectives. The transformed dendrograms, on the contrary, demonstrate a tighter clustering among the adjectives. The Average Silhouette Width Validity Index (ASWVI) [Rousseeuw, 1987] (readily available in the Matlab Statistics Toolbox and discussed in section 3.1.2) is a measure of clustering validity that indicates how appropriate the assignment of points to clusters is. It ranges from -1 to 1, with 1 showing best assignment, 0 representing average, and -1 representing inappropriate assignment. In our case the ASWVI increased after the spline ordinal transformation from 0.17 to 0.42 for the Greek data, and from -0.02 to 0.37 for the English data. A similar pattern was also observed for other relevant indices (e.g. Dunn's index [Dunn, 1974]).

This means that the application of the spline ordinal transformation has led to a higher organization of the data that in turn resulted in a clearer formulation of clusters for both linguistic groups. It is important to note here that our analytic strategy (based on preliminary factor analyses within the identified clusters) could not have been applied to the Greek data without the transformation, due to inadequate clustering.

Subsequently, the analytic strategy was applied to the original and transformed data and the results were compared. Table 5.3 shows the percentage of total and factorial variance prior to rotation that was explained by the final solution in the case of the original and spline ordinal



(a) Greek original (b) English original

(c) Greek transformed (d) English transformed

Figure 5.4: Dendrograms of the Greek (left) and English (right) adjectives before (a), (b) and after (c), (d) the spline ordinal transformation.

transformed variables. Data from the Greek original variables are not depicted because, as noted above, the deployment of the data reduction methodology was prevented due to inadequate clustering.

Table 5.3: Comparison of the amount of factor variance prior to rotation explained by different variable transformations and FA procedures (criterion used for deciding the number of factors: eigenvalues ≥ 1). Total variance is shown in bold and variance explained by each factor in parentheses. (ML: Maximum Likelihood algorithm, PAF: Principal Axis Factoring algorithm)

	Percentage of total variance	
Transform./method	**Greek**	**English**
Original/PAF	...	**77.122**
	...	(42.77, 24.54, 9.8)
Spline Ordinal/ML	**82.3**	**82**
	(36.5, 30.5, 15.2)	(48.7, 27.3, 5.9)

Table 5.3 highlights the fact that the spline ordinal transformation explained a larger proportion of total variance than the original case for the English group. Additionally, the spline ordinal transformation increased (by 8.7%) the variance explained by the first two dimensions of the English group. The higher concentration of accounted variance in the first two factors of the optimally transformed solution suggests increased correlations between the transformed variables (also evident from the dendrograms). This finding justifies the use of the optimal nonlinear approach, as the modelling of nonlinear relationships between variables led to greater explained variance by the use of fewer dimensions.

Overall, the optimal nonlinear transformation has contributed towards a more compact representation of the semantic variables (i.e. tighter clustering) which allowed the deployment of the described data reduction strategy. Additionally, FA on the transformed variables explained higher amount of total variance which was also concentrated on the first two factors compared to the untransformed case. This suggests that the transformation has indeed accounted for existing nonlinearities between the variables and has yielded a more accurate representation of the semantic space.

Maximum Likelihood algorithm for Factor Analysis

Maximum Likelihood (ML) was the preferred factor analysis algorithm (see section 3.1.3). However, the original data featured extreme positive skewness for both linguistic groups, which violated the condition of multivariate normality in the data set that is assumed by ML. Thus, the

original English group was analysed using the Principal Axis Factoring algorithm instead. The transformed data set were analysed with ML, as the spline ordinal transformation improved the conditions for its application by reducing skewness.

Two goodness-of-fit measures confirmed the validity of our FA model. The *Kaiser-Meyer-Olkin* (*KMO*) criterion[4] equalled 0.798 and 0.714 for the Greek and English-speaker dataset respectively, both of which are regarded as 'good' [Hutcheson and Sofroniou, 1999, p.225]. Bartlett's test of sphericity[5] also showed statistical significance ($p < 0.001$ for both Greek- and English-speaker datasets), revealing that the correlation matrix was significantly different from the identity matrix (i.e., the variables were not perfectly independent).

5.3.3 Intra-linguistic semantic dimensions

The transformed variables analysed with the Maximum Likelihood algorithm resulted in a 3-factor solution (eigenvalues ≥ 1) that explained the same amount of total variance (82%) in both linguistic groups (see Table 5.3). Specifically for the Greek group, the first two factors explained a similar amount of variance (36.5% and 30.5%), while the third only explained 15% of the variance. For the English group almost half of the variance (48.7%) was contained in the first factor, while the second factor explained 27.3% and the third factor only 5.9% of the total variance prior to rotation.

The emerging factors in FA are often computed as mutually orthogonal [Disley et al., 2006]. Subsequently, they are subjected to a rotation to improve the interpretability of the solution by maximising the already large factor loadings and minimising the small ones. However, in several cases, the orthogonality of the factors constitutes a strict condition and therefore can impede the interpretability of the results. Consequently, we chose to relax the requirement of factor orthogonality by employing a non-orthogonal (oblique) rotation of the initial orthogonal solution, which allows for factors to be correlated. We have used the direct oblimin method [Jennrich and Sampson, 1966], which (among others) is considered as a viable approach to the problem of oblique factor rotation [Harman, 1976].

The data reduction methodology gave the most representative verbal descriptors for this set

[4]KMO criterion assesses the sample size (i.e. cases/variables) and predicts if data are likely to factor well based on correlation and partial correlation. KMO can be calculated for individual and multiple variables and varies between 0 and 1.0. It should be 0.60 or higher to proceed with factor analysis.

[5]Bartlett's test examines the hypothesis that the correlation matrix under study is significantly different from the identity matrix, i.e. variables are not completely independent. Significance on this test confirms this hypothesis.

of sounds. These adjectives, along with their factor loadings, appear in Table 5.4 for both Greek and English groups. Factor loadings are the regression coefficients (ranging from -1 to +1) between variables and factors. Their values indicate the relative contribution that a variable makes to a factor and are crucial for the labelling and interpretation of the factors. Only descriptors with *factor loadings* $\geq$ *0.75* were considered significant in this work and will be used for factor interpretation [Tabachnick and Fidell, 2006, Comrey and Lee, 1992]. Based on the above, a proposed labelling was applied by choosing a couple of terms that we believed would better capture the essence of each semantic dimension. According to this, Factor 1 could be: *Depth-Brilliance* for Greek and *Brilliance/Sharpness* for English, Factor 2: *Roundness-Harshness* for Greek and *Roughness/Harshness* for English, and Factor 3: *Richness/Fullness* for Greek and *Thickness-Lightness* for English.

Table 5.4: Pattern matrix of the Greek and English Factor Loadings with suggested labelling after oblimin rotation. Loadings $\geq$ 0.75 are presented in bold.

	Factors					
	Greek			English		
	1 (Depth-Brill.)	**2** (Round.-Harsh.)	**3** (Rich./Full.)	**1** (Brill./Sharp.)	**2** (Rough./Harsh.)	**3** (Thick.-Light.)
Brilliant	**-0.818**	0.187	0.248	**0.989**	-0.217	-0.009
Deep	**0.913**	0.225	0.126	-0.159	-0.220	0.738
Rough	–	–	–	-0.272	**0.962**	0.084
Soft	-0.377	**0.859**	-0.088	-0.492	-0.683	-0.193
Full	0.180	-0.017	**0.835**	–	–	–
Rich	-0.321	0.115	**0.965**	–	–	–
Harsh	-0.002	**-0.934**	-0.178	0.406	**0.766**	-0.023
Rounded	0.123	**0.879**	0.196	–	–	–
Thick	**0.794**	0.160	0.364	-0.020	-0.124	**0.932**
Thin	–	–	–	0.225	0.439	-0.652
Warm	0.111	**0.906**	0.188	-0.476	-0.567	0.220
Dark	–	–	–	-0.382	0.241	0.706
Sharp	-0.488	-0.615	0.126	**0.778**	0.057	-0.035
Messy	–	–	–	-0.232	**0.882**	0.197
Light	-0.413	0.744	-0.428	-0.196	-0.212	**-0.891**
Shrill	-0.304	-0.739	0.139	0.425	0.422	-0.309
Dense	0.624	-0.078	0.541	-0.022	-0.289	**0.829**
Dull	0.620	0.489	-0.089	-0.365	-0.538	0.251
Bright	–	–	–	0.690	-0.020	-0.352

The correlation coefficients between the rotated factors together with the corresponding an-

gles ($angle = cos^{-1}(r)$) are shown in Table 5.5. The very low correlation coefficients between factors for the Greek group imply the existence of a nearly orthogonal semantic space. However, for the English group, there appears to be a mild correlation between the first and the third (58.6°) and also between the first and the second dimensions (72.8°).

Table 5.5: Inter-dimension correlations and angles.

Correlation coefficient	Greek	English
r12	0.135 (82.2°)	0.296 (72.8°)
r23	-0.009 (89.4°)	0.068 (86.1°)
r31	0.161 (80.7°)	-0.520 (58.6°)

Figure 5.5 shows the positions of the stimuli in the common factor space based on the factor scores. The presentation consists of six 2D planes resulting from the 3D Euclidean semantic timbre spaces (although dimensions are not entirely orthogonal) for both Greek and English groups. The Euclidean representation is less accurate for the English group due to its higher inter-dimensional correlation. The different symbols for each sound represent classes of musical instruments according to von Hornbostel and Sachs [1914], and the filling of the symbols represents the type of excitation (black for continuant sounds and white for impulsive sounds).

As can be noticed by visual inspection of Figure 5.5, the musical sounds' position within the common factor space (factor scores) does not provide any clear indication of possible favoured relations between the identified timbral descriptions (factor labels) and the traditionally accepted classification schemes of musical instruments. As expected, our findings further support the difficulty to identify a direct relation of musical timbre description with terms referring to broad categories of musical instruments' sounds [Campbell et al., 2006].

5.3.4 Inter-linguistic relationships

Table 5.1 presents the Spearman correlation coefficients that indicate the agreement on the use of each adjective between the two different linguistic groups. Interestingly, most of the adjectives that feature a poor inter-group correlation (e.g. *compact*, *empty*, *hollow*, *distinct* and *cold*) are also weakly correlated with the other adjectives within the linguistic groups. This is evident from the dendrograms 5.4a and 5.4b and has resulted in the removal of most of them during the data reduction phase.

A correlation analysis that resulted to the correlation matrix of Table 5.6 was subsequently

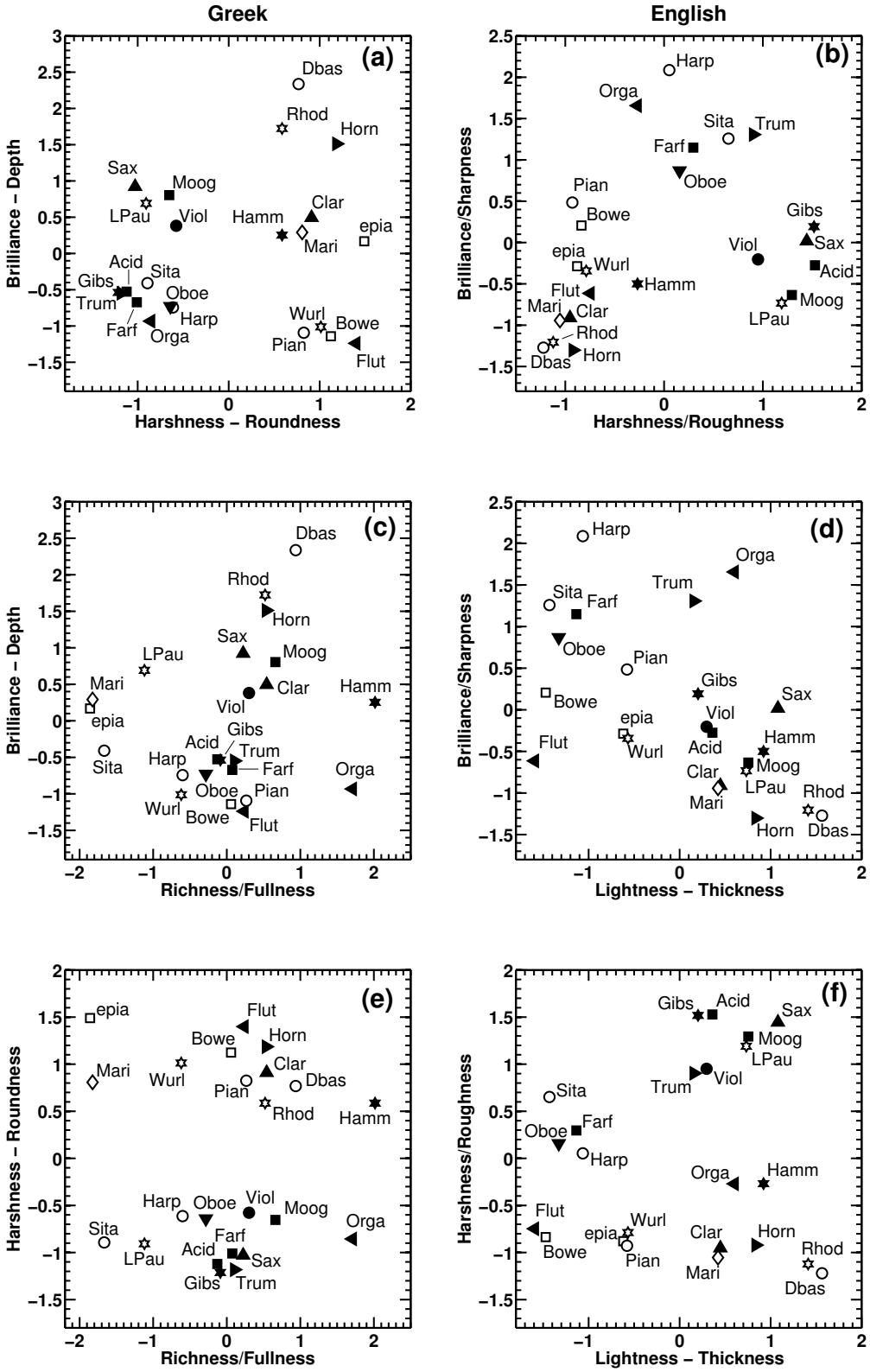


Figure 5.5: Six 2D planes of the Greek (left) and the English (right) 3D semantic timbre space. Black symbols: Continuant, white symbols: Impulsive, $\triangle$: Single reed, $\bigtriangledown$: Double reed, $\triangleleft$: Aerophone, $\triangleright$: Lip reed, $\bigcirc$: Chordophone, $\diamond$: Idiophone, $\star$: Electrophone , $\square$: Synthesiser

performed between the semantic dimensions. The Spearman correlation coefficient between the first dimensions is $\rho(21) = -0.66$, $p < 0.01$, between the second dimensions is $\rho(21) = -0.78$, $p < 0.001$ and between the third dimensions is $\rho(21) = 0.55$, $p < 0.01$. Figure 5.6 demonstrates the above by showing the scatter plots for each corresponding dimension between the two languages. While the third dimensions are only mildly correlated, the third English dimension is highly correlated with the first Greek dimension [$\rho(21) = 0.81$, $p < 0.001$] and the first English dimension shows some correlation with the second Greek dimension [$\rho(21) = -0.46$, $p < 0.05$]. This could be partly attributed to the non-negligible correlations that appear between the English dimensions presented in Table 5.5. It also shows that the terms *thickness* and *sharpness* which are included in these different dimensions are nevertheless commonly understood between the two linguistic groups. *Sharpness* as 'synonym' of *brilliance* also links that dimension with Greek *roundness-harshness*, and *thickness* strongly links the first Greek with the third English dimension. This is in agreement with the strong inter-linguistic correlations for *sharpness* and *thickness* that are evident in Table 5.1. The correlations featured across the remaining non-corresponding dimensions were not significant ($p > 0.05$).

Table 5.6: Correlation matrix between the Greek and English semantic dimensions. *: p<0.05, **: p<0.01

	1st English	**2nd English**	**3rd English**
1st Greek	-0.66**	-0.07	0.81**
2nd Greek	-0.46*	-0.78**	-0.13
3rd Greek	-0.25	-0.19	0.55**

A two-sample Kolmogorov-Smirnov test showed no significant effect of language for any dimension ($z = 0.147, p = 1.00$ between the first dimensions, $z = 0.442, p = 0.990$ between the second dimensions and $z = 0.590, p = 0.878$ between the third dimensions). The K-S test was preferred as several dimensions in each language group failed a Shapiro-Wilk normality test ($p < 0.05$).

Despite the evident similarities between the semantic spaces of the two linguistic populations, there are some differences that are also worth mentioning. The main difference concerns the terms loaded on the *brilliance* dimension for each language. The adjective *sharp* is grouped with *brilliant* in the English group but associated with *harsh* in the Greek group. This is evident both from inspection of Figure 5.4 and from Table 5.4. Additionally, it seems that *full* and *rich* form

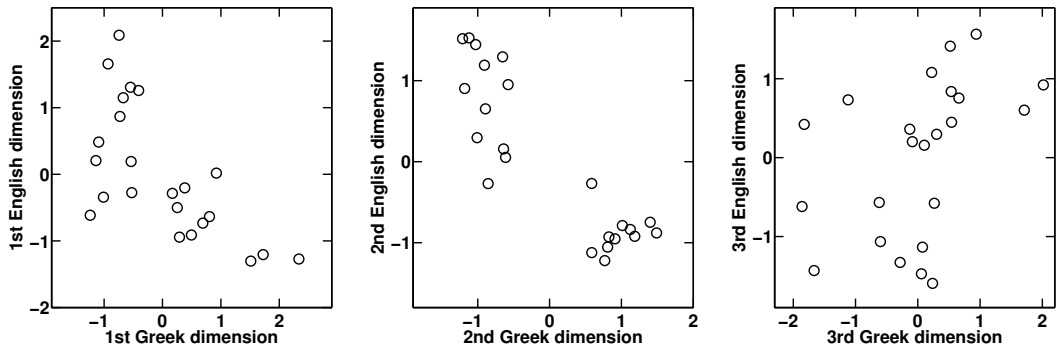


Figure 5.6: The scatter plots of the Greek and English semantic dimensions show that the 23 stimuli are similarly perceived on the corresponding dimensions. As expected from the correlation analysis, the relationship is stronger for the second dimensions and weaker for the third dimensions.

a separate group in the Greek population, whereas the same terms are more closely related to *thick*, *dense*, *deep* etc. in the English population (see Figure 5.4). As a result, *rich* and *full* form a separate factor for Greek, but *thick* and *deep* load as opposites on the *brilliance* factor. The above paragraph explains why *brilliance* dimension is enriched with unrelated terms for each of the two linguistic groups.

The extra terms provided by the listeners (see table 5.7) generally fall into seven conceptual categories as grouped by the author for both populations:

1) *properties of source* (wooden, glassy, synthetic, etc.)

2) *temporal evolution* (static, energetic, constant, etc.)

3) *emotional terms* (sinister, oppressive, suave, etc.)

4) *technical terms* (spectral, phasey, sinewave, etc.)

5) *sense of sight* (blurred, smoky, transparent, etc.)

6) *sense of touch* (raspy, gentle, blunt, etc.)

7) *size of object* (large, majestic, heavy, etc.)

These categories appeared to be more evident in the English group because of the larger number of extra terms given (117 extra terms in English compared to the 31 extra terms in Greek). The lack of terms in the last three categories can be explained by the fact that they were already well represented in the provided adjectives. The three largest categories in both linguistic groups were *properties of source*, *temporal evolution* and *emotional terms*. The only predefined descriptor belonging to one of these three categories was *metallic*.

5.4 Discussion

The analysis presented in the previous section has identified three semantic dimensions that explain more than 80% of the variance in the descriptive data. These dimensions show high independence for the Greek group while the inter-dimensional correlation is moderate between some dimensions for the English participants.

The application of an optimal nonlinear transformation supported the existence of nonlinearities by providing a more compact representation of the data and explaining more variance in the first two dimensions for both groups. It can be argued that the transformation did not affect the qualitative interpretation of the semantic dimensions. However, the value of this approach lies in the output of a more accurate representation of the positions of the sound stimuli within the identified semantic timbre space. This is particularly significant for the search of acoustic correlates and for investigating the association of semantic with perceptual spaces.

As mentioned in section 2.5, there exists evidence that language affects the way people think about objects. Contrary to this, our work was partly motivated by an intuitive assumption that timbre semantics could feature a general agreement across languages. Although this assumption was not subjected to a thorough hypothesis-inference scrutiny (which would require careful control of several additional parameters and factors), we demonstrated that the three pairs of semantic dimensions for the two linguistic groups share common conceptual properties. This exploratory approach, supported by some preliminary inferential tests (K-S and Spearman correlation), provides strong indication that despite the differences in the use of individual descriptors, there exists a similar semantic space for timbre between these two languages, at least for this stimulus set. In addition, it justifies further investigation of hypotheses regarding the universality of timbre semantics.

Therefore, we will propose an empirical labelling to express the common concept for each

Table 5.7: Collection of descriptors from free verbalization. The number in parentheses represents the number of different participants that have used the term. The Greek terms (appearing in parentheses below the English equivalent) were translated into English by the authors.

	Semantic Category						
Group	**source properties**	**temp. evolution**	**emotional terms**	**sight**	**touch**	**size**	**technical**
English	wooden (6), elastic glassy (4), scraping synthetic (4), wet percussive (2) breathy (3), plastic electronic (2), real buzzy, brassy, bassy natural, twangy (2) reedy (3), steely airy, unnatural (2) pianolike, organlike desiccated, ethereal artificial (3), farty resonant (2), sterile organic, futuristic alien, pure (3), jingly complex (5), distant muffled, tinny (2)	wavey (4), flat (5) energetic, rising constant, fluctuating unstable, oscillating stable, vibrating (4) continuous, static pulsating (2) phase-beating wobbly, cycling (3) throbbing, varied unsettled, evolving spinning, consistent moving (2), bouncy	sinister, confusing (2) oppressive, trivial suave, intriguing relentless, boring interesting, ugly keen, unattractive annoying, brittle disorientating, neutral unpleasant (2), sickly attractive, harmless	blurred focused transparent diffused fuzzy smoky golden indistinct	raspy (2) gentle blunt textured piercing (2) penetrating grating wooly	large majestic heavy full bodied forceful substantive limited superficial shallow 1-dimensional 3-dimensional	spectral (2) phasey sinewave morphing distorted overtoney vibrating (4) resonant (2) harmonic
Greek	spacey (3) (διαστημικός) muffled (μουκωμένος) Indian (Ινδικός) fake (4) (ψεύτικος) electronic (3) (ηλεκτρονικός) noisy (θορυβώδης)	abrupt (απότομος) discontinuous (ασυνεχής) vibrated (βιμπράτο) unstable (3) (ασταθής)	sweet (4), unsure (γλυκός), (αβέβαιος) hesitant , funny (δισταχτικός), (αστείος) relaxing , psychedelic (χαλαρωτικός), (ψυχαιδελικός) befooling , emetic (κοροϊδευτικός), (εμετικός) dizzying , hypotonic (ζαλιστικός), (υποτονικός) nice, annoying (2) (συμπαθητικός), (ενοχλητικός) hair-raising (ανατριχιαστικός) lacking vividness (χωρίς ζωντάνια)	transparent (διάφανος) indistinct (δυσδιάκριτος)	squeaky (τσιριχτός)	dynamic (δυναμικός) intense (έντονος) exaggerated (υπερβολικός)	echo

of the semantic dimensions. The dimension that shows the strongest agreement between the two groups is the one that describes whether a sound is perceived as smooth-and-round or rough-and-harsh. As these adjectives originate from tactile quality description we suggest the label *texture* for this dimension. The first dimensions for both linguistic groups have the adjective brilliant in common. This is a metaphor that comes from the domain of vision, we therefore suggest the label *luminance* for the description of this dimension. Finally, the third dimensions in both groups describe whether a sound is perceived as thick-dense-rich-and-full or light. We suggest *mass* as an appropriate general semantic label for this dimension.

These results seem to support Lichte [1941] who concluded that: "... complex tones have, in addition to pitch and loudness, at least three attributes. These are brightness, roughness, and one tentatively labelled fullness. The first two are probably more basic than the third". There also seems to be some agreement regarding the number and naming of dimensions with some earlier studies [von Bismarck, 1974a, Pratt and Doak, 1976, Moravec and Štěpánek, 2003, Štěpánek, 2006, Alluri and Toiviainen, 2010]. Taken as a whole, there appears evidence that the major semantic dimensions of timbre are language-independent.

In agreement with these studies, the boundaries between semantic dimensions are not always clearly defined. *Luminance* and *mass* dimensions are correlated with each other, particularly for the English group. Sounds that are described as *light* are more likely to also be described as *brilliant*, while sounds described as *thick* or *dense* are also described as *less brilliant*. Additionally, we provide some evidence that *luminance* is conceptually related to *texture* in the English language as suggested by the fact that *sharpness* (a term that is positioned in the *texture* cluster in Greek dendrograms 5.4a and 5.4c) is highly loaded (0.778) on the *luminance* dimension. This last finding is not unexpected as Štěpánek [2006] has supported that *sharpness* is an auditory attribute that lies between *luminance* and *texture* (i.e., a sound object featuring both high *luminance* and high *texture* is described as *sharp*). However, the interpretation of specific differences (mainly some unrelated terms loaded on the *luminance* dimension) between the semantic dimensions of the two language populations would require a linguistic analysis which, although interesting per se, lies beyond the scope of this work.

The acquisition of extra terms from spontaneous descriptions suggests that future researchers on timbre semantics should consider including terms that belong to one additional semantic category: *temporal evolution*. Although the number of terms acquired for description of the *prop-*

erties of source and *emotions* is also considerably large, they should probably be avoided when studying the semantic description of sound impressions [Wake and Asahi, 1998].

Finally, while it has been shown that same-family instruments tend to occupy similar regions in perceptual spaces resulting from pairwise dissimilarity ratings [Giordano and McAdams, 2010], this can not be supported by the semantic space structure of this work. As a possible explanation, it can be assumed that while perceptual spaces resulting from cognitive dissimilarity ratings and MDS analyses represent both sensory and semantically meaningful factors, verbal attribute studies can only capture the semantically charged portion of the MDS spaces. Consequently, the comparison of these semantic spaces with perceptual spaces resulting from a pairwise dissimilarity experiment using the same stimuli could be proven useful in testing the above hypothesis.

5.5 Acoustic correlates of semantic dimensions

A large set of low-level features (see subsection 3.2.2) was extracted from the experimental sound set as an initial attempt to identify acoustic correlates for the semantic dimensions that resulted from Factor Analysis. Table 5.8 presents only the audio descriptors that were found to be perceptually significant and section 3.2.2 presents all the descriptors that were initially extracted along with their formulas. The selection of acoustic features was based on the existing literature [e.g. Peeters, 2004, Peeters et al., 2011] and they were calculated using the Spectral Modeling Synthesis (SMS) Matlab platform [Amatriain et al., 2002]. A short description of the SMS model is provided in subsection 3.2.1. The window length applied was 4096 samples ($f_S =$ 44.1kHz) with an overlapping factor of 87.5%, the zero padding factor was 2 and 50 harmonic partials were extracted for all sounds. A variation of some basic features was also extracted using the instantaneous specific loudness of the ERB bands as calculated by Moore's loudness model [Moore et al., 1997] instead of the amplitude of the harmonics or the FFT bins. In order to avoid the effect of the low signal-to-noise ratio (SNR) in the tail of the release (especially for percussive sounds) on the feature calculation, we cropped all our sounds to the point where the SNR dropped below 25 dB[6]. The energy of the noise was calculated as the average energy of the last 10 frames of the signal (window: 1024, hop size: 128). Moreover, the sounds were also cropped in the beginning at the point where the local energy ratio remained above 1 dB so

[6] An SNR value above 25 dB is usually regarded acceptable for many applications (e.g. image processing, wireless communications etc.) [Stremler, 1990].

as to discard the initial silent gap before the onset. Special attention has been paid to avoid the introduction of any artifacts from this processing procedure.

The problem of strongly correlated clusters of acoustic features needed to be addressed before proceeding to correlation analysis with the semantic dimensions. One approach would be to consider an acoustic feature as significantly associated with a dependent variable only when both their correlation and partial correlation were significant [Giordano et al., 2012]. However, while this approach avoids data reduction methods, it discards variance that is common between features. Thus, an exploitation of Principal Components Analysis (PCA) was favoured similarly to Alluri and Toiviainen [2010], Giordano et al. [2010] and Peeters et al. [2011]. To reduce high multicollinearity within the variable (feature) set, we initially inspected the Spearman coefficient correlation matrix and discarded strongly correlated features [$\rho(21) \geq 0.8$]. We then rank-ordered the features and applied PCA to the reduced data set. Inspection of the anti-image correlation matrix[7] diagonal led to further removal of features whose individual Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) was less than 0.5 so as to achieve an acceptable overall KMO. The final solution consisted of 4 components with eigenvalues ≥ 1 (KMO = 0.673, Bartlett's test of sphericity $p < 0.001$) that explained 83.3% of the total variance. Table 5.9 shows the loadings of the features on the 4 components after orthogonal Varimax rotation. The components are labelled based on the acoustic correlates that are highly loaded on each one.

Features like the *normalised harmonic spectral centroid* (SC_norm), *tristimulus 3* (T3) [Pollard and Jansson, 1982], *SC_loud_cor* (corrected version of the spectral centroid in order to remove the influence of F_0, for an example see Marozeau and de Cheveignè [2007]) and *harmonic spectral spread* (Spread) all represent spectral structure (i.e. distribution of energy among harmonic partials) rather than spectral content. Therefore, the first component is labelled: *energy distribution of harmonic partials*. The second component is related to *spectrotemporal* characteristics such as *noisiness*, *harmonic spectral flux* (Flux) and the *standard deviation of the harmonic spectral centroid* (SC_std). The third component is represented by both *spectral centroid variation (SC_var_loud)* calculated from Moore's specific loudness [Moore et al., 1997] and *inharmonicity*. Finally, the fourth component is related to a temporal characteristic like *the logarithm of the attack time* (Log_At_time) and a spectrotemporal one like the temporal variation of the first nine harmonics (*Mean coefficient of variation, MCV,* Kendall and Carterette, 1993b).

[7]The anti-image correlation matrix contains measures of sampling adequacy for each variable along the diagonal and the negatives of the partial correlation on the off-diagonals.

Table 5.8: Abbreviations and definitions of the significant audio features.

Category	Feature	Abbreviation	Explanation
Spectral Content	Harmonic Spectral Centroid	SC	Barycenter of the harmonic spectrum [Peeters et al., 2011]
	Spectral Centroid (loudness model)	SC_loud	SC of the specific loudness [Moore et al., 1997]
Energy distribution of harmonic partials	Normalised Harmonic Spectral Centroid	SC_norm	Normalised barycenter of the harmonic spectrum
	Tristimulus 1, 2, and 3	T1, T2, T3	Relative amplitudes of the 1st, the 2nd to 4th and the 5th to the rest harmonics [Pollard and Jansson, 1982]
	Harmonic Spectral Spread	Spread	Spread of the harmonic spectrum around its mean value [Peeters et al., 2011]
	SC (loudness model) corrected	SC_loud_cor	SC of the specific loudness corrected for F_0 (Moore et al., Marozeau and de Cheveignè)
Spectrotemporal	Harmonic Spectral Flux (or variation)	Flux	Amount of variation of the harmonic spectrum over time [Krimphoff, 1993]
	Mean Coefficient of Variation	MCV	Variation of the first 9 harmonics over time [Kendall and Carterette, 1993b]
	SC standard deviation	SC_std	SC standard deviation over time
	SC variation	SC_var	SC_std/SC_mean [Krimphoff, 1993]
	SC variation (loudness)	SC_var_loud	SC variation of the specific loudness
	Noisiness	Noisiness	Ratio of the noise energy to the total energy [Peeters et al., 2011]
Spectral fine structure	Harmonic Spectral Irregularity	Sp_Irreg	Measure of the harmonic spectrum fine structure [Kendall and Carterette, 1996]
	Odd Even Ratio	OER	Ratio of the energy contained in odd versus even harmonics [Peeters et al., 2011]
Harmonic series	Inharmonicity	Inharmonicity	Peeters et al. [2011]
Temporal	Log of attack time	Log_At_time	Logarithm of the rise time [Peeters et al., 2011]
	Temporal Centroid	TC	Barycenter of the energy envelope Peeters et al. [2011]
	Normalised Temporal Centroid	TC_norm	TC/duration

Table 5.9: Loadings of the audio features on the first 4 principal components as a result of PCA with Varimax rotation. Loadings ≥ 0.75 are presented in bold and used for labelling the components.

	Component			
	1	**2**	**3**	**4**
	(Energy distribution of harm. partials)	(Spectrotemporal)	(Spectrotemporal, Inharmonicity)	(Temporal, Spectrotemporal)
SC_norm	**0.955**	-.030	0.170	-0.012
T3	**0.931**	-0.127	0.110	0.054
SC_loud_cor	**0.876**	-0.248	-0.316	0.062
SC_loud	**0.794**	-0.201	-0.488	0.052
Spread	**0.785**	-0.107	-0.419	-0.167
T2	-0.734	0.066	-0.473	0.203
Noisiness	0.047	**0.909**	0.254	-0.209
Flux	-0.199	**0.875**	0.058	-0.016
SC_std	-0.342	**0.823**	0.176	-0.39
SC_var_loud	-0.138	0.391	**0.790**	-0.132
Inharmonicity	0.272	0.301	**0.789**	-0.140
OER	-0.382	-0.41	0.650	-.336
Log_At_time	0.006	0.055	-0.235	**0.829**
MCV	-0.223	-0.445	-0.016	**0.761**
TC_norm	0.149	-0.574	-0.211	0.576

Table 5.10 presents the Spearman correlations coefficients between the mutually orthogonal components and the semantic dimensions (factor scores) for both linguistic groups. F_0 has been also considered in the correlation analysis in order to reveal its potential influence on the semantic dimensions.

5.5.1 Greek intra-group results

The *Luminance (Depth/Thickness-Brilliance)* dimension shows significant positive correlation [$\rho(21) = 0.68$, $p < 0.01$] with the third principal component (SC variation and inharmonicity) and is also influenced by the fundamental frequency [$\rho(21) = -0.58$, $p < 0.01$]. The *Texture (Roundness-Harshness)* dimension shows a strong negative correlation [$\rho(21) = -0.75$, $p < 0.001$] with the first component which represents the energy distribution of harmonic partials. The *Mass (Richness/Fullness)* dimension does not exhibit strong correlations with any of the principal components.

5.5.2 English intra-group results

The *Luminance (Brilliance/Sharpness)* dimension is correlated with the energy distribution of harmonic partials [$\rho(21) = 0.61$, $p < 0.01$] and is weakly correlated [$\rho(21) = -0.50$, $p < 0.05$] with the third principal component (SC variation and inharmonicity). The *Texture (Harshness/Roughness)* dimension exhibits strong correlation [$\rho(21) = 0.74, p < 0.001$] with the energy distribution of harmonic partials. Finally, the *Mass (Thickness-Lightness)* dimension features strong correlation [$\rho(21) = 0.7, p < 0.001$] with the third principal component (SC variation and inharmonicity) and is also heavily influenced by the fundamental frequency [$\rho(21) = -0.76, p < 0.001$].

Table 5.10: Spearman correlation coefficients between semantic dimensions, the 4 principal components of the audio feature set and F_0 (***: p<0.001, **: p<0.01, *: p<0.05). Coefficients that feature significance levels above p<0.01 are highlighted in bold.

		Energy distribution of harmonic partials	Spectrotemporal	Spectrotemporal, Inharmonicity	Temporal, Spectrotemporal	F_0
Greek	Depth/Thickness-Brilliance	-0.119	-0.260	**0.681****	0.155	**-0.581****
	Roundness-Harshness	**-0.754*****	0.114	-0.181	0.037	0.436*
	Richness/Fullness	-0.028	-0.186	0.032	0.440*	-0.230
English	Brilliance/Sharpness	**0.615****	0.199	-0.503*	0.065	0.276
	Harshness/Roughness	**0.737*****	-0.132	0.011	-0.044	-0.178
	Thickness-Lightness	-0.084	-0.183	**0.704*****	0.218	**-0.756*****

5.5.3 Inter-linguistic comparison and discussion

The second part of this work examined possible relationships between the uncovered semantic dimensions and acoustic characteristics of the sound stimuli. As shown in Table 5.10, the conceptually related semantic dimensions between the two languages did not always correspond to the same acoustic dimensions. The most important factor for the auditory perception of *texture* seems to be the *energy distribution of harmonic partials*. The correlations for both linguistic groups indicate that sounds with stronger high partials are more likely to be characterised as *rough* or *harsh* and the opposite as *round* or *soft*. This appears to support Faure et al. [1996], Howard and Tyrrell [1997], Barthet et al. [2010a] and Barthet et al. [2011] who have generally associated higher spectral centroid values with *roughness* and *shrillness* and lower spectral centroid values with *softness*.

Luminance featured significant correlation with spectral structure only in the English group,

but there is some evidence that the amount of inharmonicity influences auditory brilliance (i.e., more inharmonic sounds are perceived as less brilliant) in both groups. Additionally, sounds with a stronger spectral centroid fluctuation are also more likely to be perceived as less brilliant. There is some evidence that fundamental frequency is positively correlated with brilliance in the Greek group. The findings concerning *luminance* and *texture* seem to support Schubert and Wolfe [2006] whose empirical study has proposed that *simple SC* is a better correlate for perceptual brightness than the *normalised SC*. In other words, these results suggest that the distribution of energy, as expressed by the normalised SC, seems to be a better correlate of texture whereas spectral content (also related with F_0) might predict luminance more efficiently.

Mass did not correlate significantly with any component in the Greek group. On the contrary, it exhibited two strong correlations in the English group. These correlations suggested that sounds with higher F_0 were perceived as lighter and also that auditory thickness and density increased with inharmonicity and with fluctuation of the spectral centroid. The latter is in some agreement with Terasawa's definition of density [Terasawa, 2009] as "the fluctuation of instantaneous intensity of a particular sound, both in terms of rapidity of change and degree of differentiation between sequential instantaneous intensities".

Overall, the combination of the Greek and English group findings suggest that *texture* is evidently affected by the energy distribution of harmonic partials. The picture is not so clear for *luminance* and *mass* and future research on their acoustic correlates is mandated. However, there are indications that auditory thickness is enhanced by inharmonicity and SC fluctuation, whereas auditory brilliance is decreased. The influence of F_0 was more evident in the English group's perception of mass and less evident in the Greek group's perception of luminance indicating that the effect of F_0 on timbre semantics needs to be further investigated.

5.6 Conclusion

This study investigated the underlying structure of musical timbre semantics through an analysis of verbal description of different timbres. Factor and cluster analyses were performed on semantic descriptors that were obtained from two linguistic groups (Greek and English) for musical instrument tones. The salient semantic dimensions for timbre description were identified and compared between the two linguistic groups. A correlation analysis between extracted acoustic descriptors and semantic dimensions indicated the prominent acoustic correlates. The major

contributions of this work can be summarised as follows:

(1) The statistical analysis results suggested the existence of nonlinear relationships between the semantic variables. An optimal nonlinear transformation applied to the raw data accounted for such nonlinearities between the variables and resulted in a more efficient modelling of their underlying structure. This means that linear modelling of such data should be undertaken with care.

(2) While there did not seem to be consensus in the use of every descriptive adjective between the two linguistic groups (see Table 5.1), the three identified semantic dimensions exhibited a high degree of similarity. These common semantic dimensions could be labelled as *luminance*, *texture* and *mass*. This is an indication of language-independent description of musical timbre, at least between English and Greek.

(3) The strongest acoustic correlates identified for both linguistic groups were the following: i) the energy distribution of harmonic partials was associated with *texture*, ii) inharmonicity and variation of the SC were positively correlated with *thickness* and negatively correlated with *brilliance*, iii) F_0 affected English *mass* negatively and Greek *luminance* positively.

The following chapter of this thesis examines the relationship between semantics and perception of musical timbre through comparison of this descriptive approach to a pairwise dissimilarity rating approach and multidimensional scaling (MDS) analysis.

Chapter 6

Semantics vs perception

6.1 Introduction

The findings of chapter 5 suggest that musical timbre semantics feature strong similarities across languages. This work focused merely on *iconic musical meaning* [Koelsch, 2011], that is timbral descriptions associated with sounds and qualities of objects or qualities of abstract concepts. The next step will be to examine the relationship of the universal timbral semantics with perception.

As stated in chapter 2, the close relationship between verbally described timbral dissimilarities and numerical dissimilarity ratings found by Samoylenko et al. [1996] has not been confirmed at the level of underlying perceptual dimensions. This chapter will examine whether the salient semantic dimensions revealed previously (i.e. *luminance, texture and mass*) correspond to underlying perceptual dimensions that come from non-verbal assessment of timbre. To this end, the timbre spaces that resulted from two different relational measures experiments (a VAME listening test presented in chapter 5 and a pairwise dissimilarity rating experiment) were compared. Unlike other related studies [e.g. Faure et al., 1996, Elliott et al., 2012] the participants in our work were different for each separate listening test.

Since the two inter-language semantic spaces featured many common elements, this chapter will examine the relationship between the English semantic timbre space and the perceptual timbre space of the same stimuli (obtained through non-metric multidimensional scaling analysis of the dissimilarities). A potentially strong relationship between the two timbre spaces, acquired through distinct experimental procedures, would highlight the value of musical timbre semantics,

a fact that could be further utilised for intuitive sound processing applications.

6.2 Method

The first timbre space of the comparison resulted from multidimensional scaling (MDS) analysis that was applied to pairwise dissimilarity ratings. The second space was the outcome of Factor Analysis applied to the data of a VAME listening test undertaken by 41 native English speakers as described in chapter 5. The VAME listening test consisted of 23 stimuli (discussed in subsection 5.2.1) while one additional cello tone was included in the pairwise dissimilarity test.

In the pairwise dissimilarity listening test participants were asked to compare all the pairs among the 24 sound stimuli. Therefore, they rated the perceptual distances of 300 pairs (*same-sound* pairs included) by freely inserting a number of their choice for each pair with 0 indicating an identical pair. The ratings were then normalised for each listener. This approach was preferred over the typical slider with fixed scale as it offered flexibility of rating especially for the highly dissimilar pairs. The interface was built in Matlab and is shown in Figure 6.1.

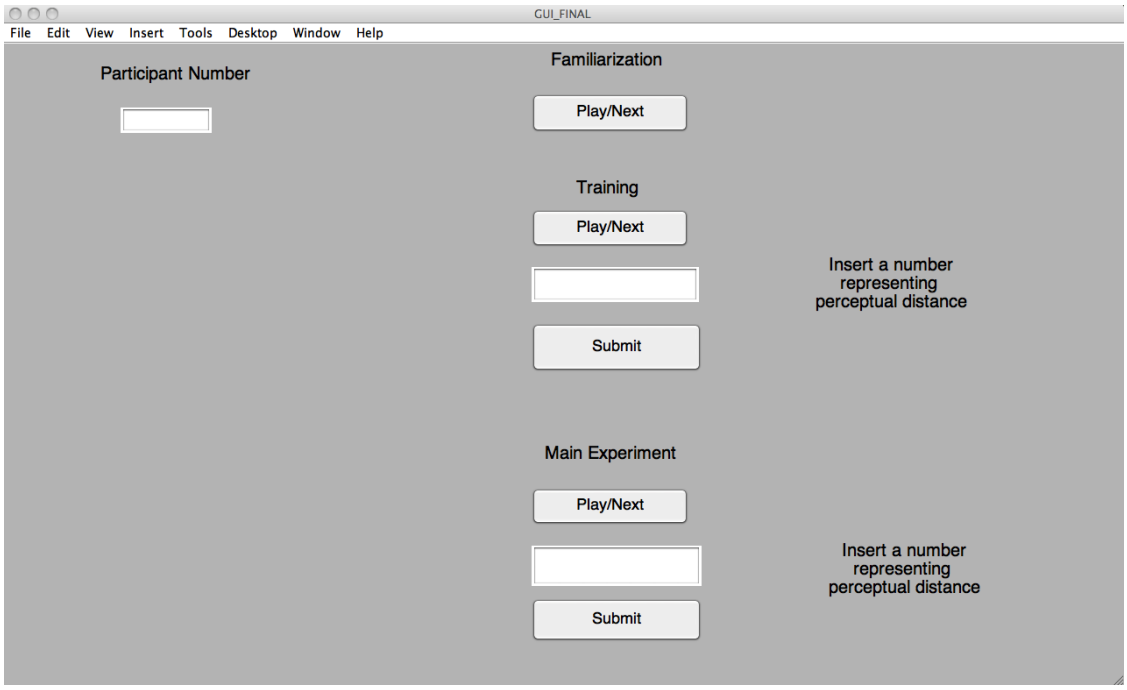


Figure 6.1: The Matlab interface of the pairwise dissimilarity experiment featured a familiarisation and a training stage together with the main experiment stage.

6.2.1 Stimuli and Apparatus

The stimulus set was identical to what was described in 5.2.1 with an additional cello tone from the MUMS library [Opolko and Wapnick, 2006] at A3 (220 Hz).

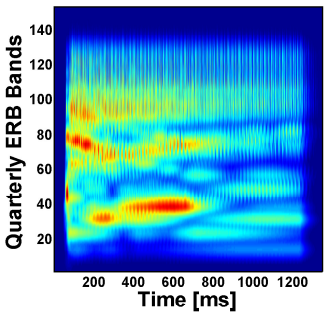
In contrast to the VAME test where sounds varied in both duration (from 3 to 8 secs) and pitch, these two variables needed to be equalised as much as possible for the pairwise dissimilarity test. To this end, only the first 1.3 seconds of each sound were retained with an exponential fade out (in linear amplitude scale) applied to the last 113 msecs (i.e 5000 samples)[1]. Furthermore, the 5 sound samples at G3 and A3# were pitch shifted to A3 so that the whole sound set consisted of merely chroma class 'A' (ranging from A1 to A4). Krumhansl and Iverson [1992] have stated that even though pitch and timbre are not perceived independently this does not imply that a comparison of timbres with different pitches is impossible. Marozeau et al. [2003] and Marozeau and de Cheveignè [2007] have also shown that listeners were able to ignore pitch differences and focus merely on timbre for a range up to at least 1.5 octave. The inter-stimulus interval (ITS) was set to 0.5 secs. The sound samples were loudness equalised in an informal listening test within the research team. The resulting RMS playback level was measured between 65 and 75 dB SPL (A-weighted). All the participants found that level comfortable for all stimuli and reported that loudness was perceived as being constant across stimuli in a subsequent questionnaire based evaluation. The spectrograms of the 24 sounds according to Moore's loudness model [Moore et al., 1997] are shown in Figure 6.2.

The listening test was conducted under controlled conditions in acoustically isolated listening rooms. Sound stimuli were presented through the use of a laptop computer with an M-Audio (Fast Track Pro USB) external audio interface and a pair of Sennheiser HD60 ovation circumaural headphones.

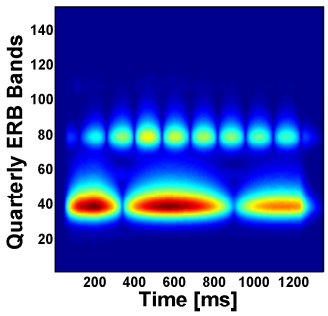
6.2.2 Participants

Thirty five listeners (aged 19-50, mean age 24, 19 female) participated in the listening test. None of the participants reported any hearing loss or absolute pitch and all of them had been practising music for 13.2 years on average, ranging from 6 to 25. The absence of absolute pitch from the group of our participants was a prerequisite as such a condition could affect the results due to

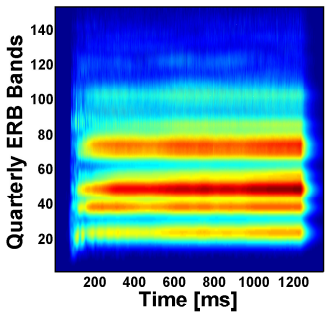
[1]The shortened equal duration could not exceed the minimum duration of the impulsive sounds in the set. Therefore, 1.3 seconds was a value that did not violate this condition while being long enough to preserve the timbral quality of the sounds. The exponential fade out of 5000 samples was selected to be short, as a longer duration would impose an identical release to all of the sounds in the set.



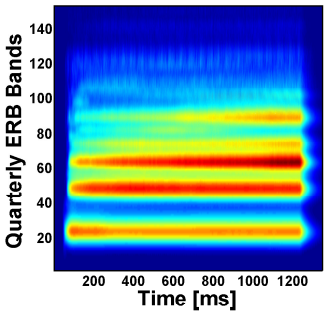
(a) Acid



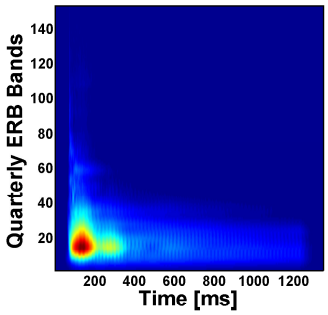
(b) Bowedpad



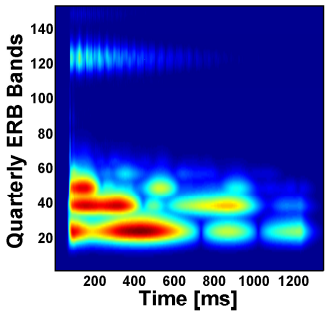
(c) Cello



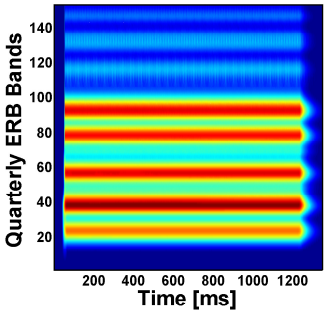
(d) Clarinet



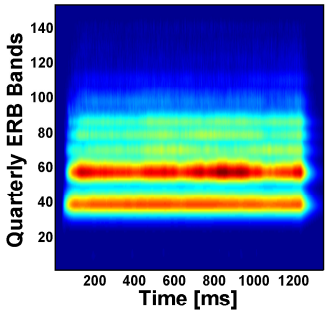
(e) Double Bass



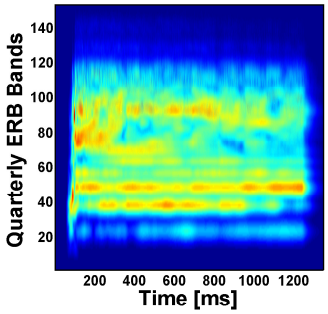
(f) epiano



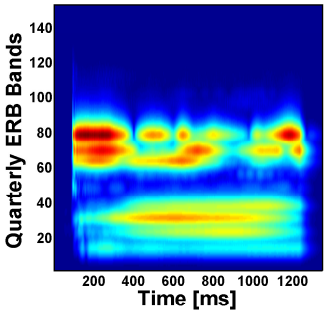
(g) Farfisa



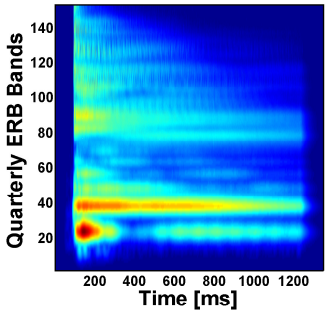
(h) Flute



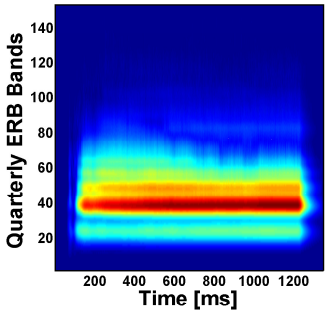
(i) Gibson



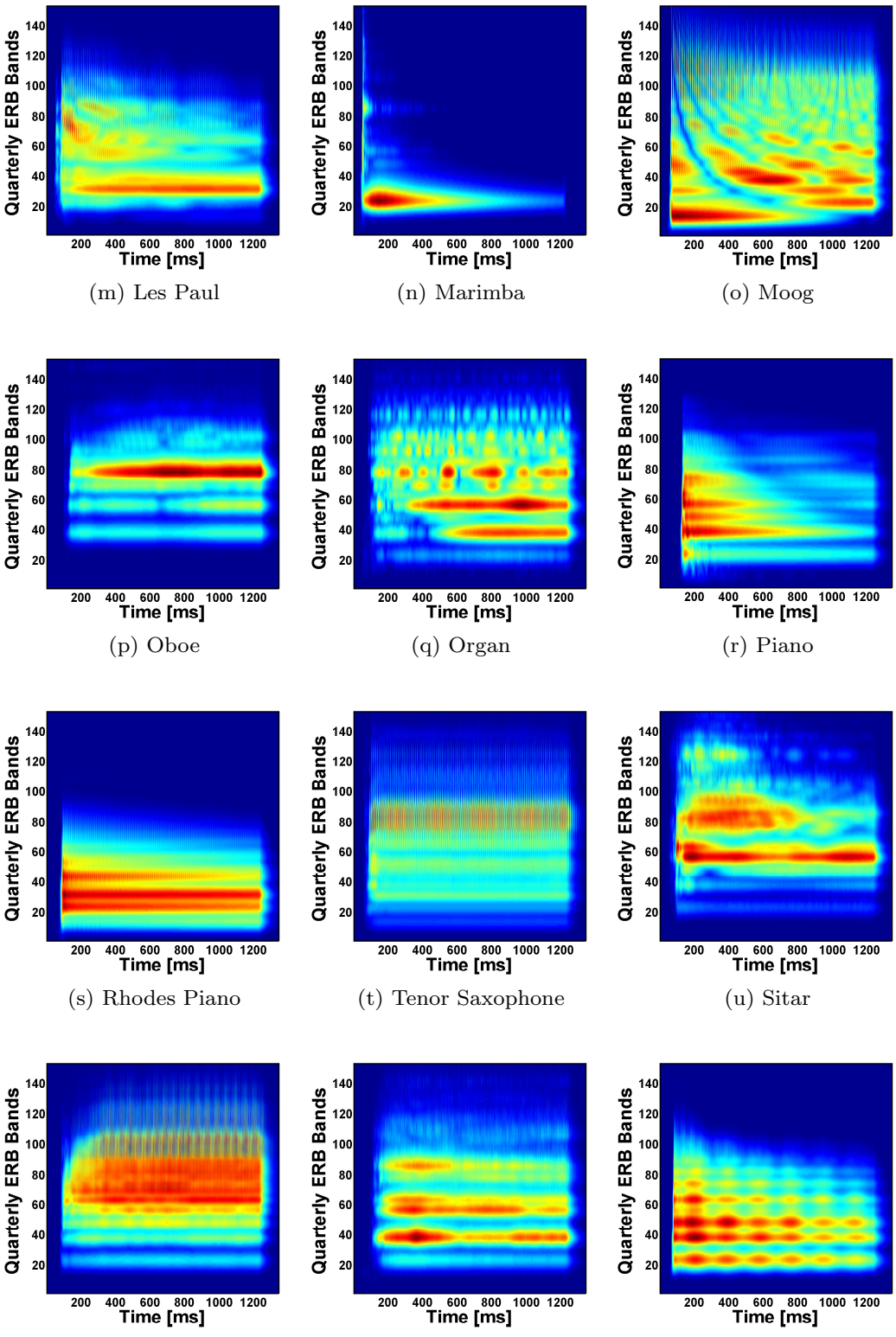
(j) Hammond



(k) Harpsichord



(l) French Horn



(m) Les Paul (n) Marimba (o) Moog

(p) Oboe (q) Organ (r) Piano

(s) Rhodes Piano (t) Tenor Saxophone (u) Sitar

(v) Trumpet (w) Violin (x) Wurlitzer

Figure 6.2: Spectrograms of the 24 stimuli used for the pairwise dissimilarity experiment. The spectrograms resulted from Moore's loudness model [Moore et al., 1997]. Y axis represents frequency by 153 quarterly ERB bands and x axis represents time in milliseconds.

pitch variation within the stimulus set. Participants were mostly students in the Department of Music Studies of the Aristotle University of Thessaloniki and a few research students from the Centre for Digital Music at Queen Mary University of London.

6.2.3 Procedure

The listeners became familiar with the timbral range of the experiment during an initial random presentation of the stimulus set. This was followed by a brief training stage where listeners rated five selected pairs of stimuli. For the main part of the experiment participants were allowed to listen to each pair of sounds as many times as needed prior to submitting their rating. The pairs were presented in random order and listeners were advised to base their ratings merely on timbral differences ignoring differences in pitch and to maintain a consistent rating strategy throughout the experiment. Participants were prompted to take one break at the completion of the first third and a second one at the completion of the second third of the overall experiment. The overall listening test procedure, including instructions and breaks, lasted around one hour for most of the participants.

The above procedure was repeated twice by each participant in two successive days. The first take was treated as a practice run and was discarded, while the second take was treated as the main listening experiment whose results were further analysed.

6.2.4 Non-metric Multidimensional Scaling

The non-metric MDS type [Shepard, 1962b, Kruskal, 1964b] that was preferred for this work makes only ordinal assumptions about the data and has been proven robust to the presence of monotonic transformations or random error in the data [Shepard, 1966, Young, 1970]. A weighted Euclidean PROXSCAL algorithm (see subsection 3.1.1) was utilised through the SPSS statistics software.

6.3 Analysis and Results

6.3.1 Non-metric MDS analysis

The Cronbach's Alpha (see subsection 3.1.5) among participants exceeded 0.9 indicating high inter-participant reliability. The ratings of the thirty five participants were analysed through non-metric weighted MDS. Table 6.1 shows two measures-of-fit (S-Stress and DAF) described in

subsection 3.1.1 along with their improvement for each added dimension. The optimal dimensionality was judged to be three as the improvement of the measures-of-fit from a 3D to a 4D space solution was minimal. The measures-of-fit for the non-metric approach were better than those of the metric approach for the same dimensionality. Furthermore, all participants have been attributed very similar weights for all the dimensions meaning that their judgements were based on the same criteria.

Table 6.1: Measures-of-fit and their improvement for different MDS dimensionalities.

Dimensionality	**S-Stress**	Improv.	**DAF**	Improv.
1D	0.3410	–	0.8130	–
2D	0.1950	0.1460	0.9176	0.1046
3D	0.1217	0.0733	0.9550	0.0374
4D	0.0951	0.0266	0.9682	0.0132

6.3.2 Comparison of the perceptual MDS space with the English semantic space

A clockwise rotation relative to the x, y and z axes with a step of five degrees (5°) was then applied to the resulting 3D matrix of coordinates. The rotated versions of the MDS space were subsequently compared to the semantic space[2]. The correlation matrix between the dimensions of the MDS space and the ones of the semantic space was then calculated. The sum of the maximum Spearman correlation coefficient for each MDS dimension was used as a criterion of optimal fit between each of the two compared timbre spaces. The three rotation angles relative to the x, y and z axes that maximised this sum were $\theta = 105°$, $\phi = 185°, \psi = 25°$. Table 6.2 shows the Spearman correlation coefficients between the optimally rotated MDS space and the semantic space.

As shown in Table 6.2, *luminance* as well as *mass* dimensions are both correlated with the second MDS space dimension [$\rho(21) = -0.68$ and $\rho(21) = 0.81$ respectively, $p < 0.001$]. This is not unexpected since these two perceptual dimensions feature some mild correlation (see chapter 5). The *texture* dimension appears to be correlated with both the first and the third MDS dimensions [$\rho(21) = -0.7$ $p < 0.001$ and $\rho(21) = -0.62$ $p < 0.01$ respectively]. Multiple regression models of both directions (i.e. with the semantic and perceptual dimensions alternating in the roles of dependent and predictor variables) have been examined. The only case in which the

[2]The cello tone was removed from the MDS space to enable direct comparison with the semantic space of chapter 5.

Table 6.2: Spearman correlation coefficients between the English semantic space and the optimally rotated MDS space. The labelling of the dimensions is according to chapter 5 (***: p<0.001)

Dimensions	**1st MDS**	**2nd MDS**	**3rd MDS**
Luminance	-0.37	-0.68***	-0.37
Texture	-0.70***	0.00	-0.62**
Mass	-0.08	0.81***	-0.01

stepwise multiple regression analysis provided a valid model was when dimensions 1 and 3 were used as independent variables and *texture* as dependent variable. The results (shown in the regression equation 6.1 and Table 6.3) are merely used for the representation of *texture* within the timbre space rather than for suggesting any causal relationships between the variables.

$$texture = -0.63 \times D1 - 0.56 \times D3 + 0.007 + err \tag{6.1}$$

This shows that dimensions 1 and 3 are almost equally important in determining the position on *texture* dimension. The summary of the multiple regression model which accounts for 84% of *texture* variance appears in Table 6.3.

Table 6.3: Stepwise multiple regression with *texture* as dependent variable and dimensions 1 and 3 as predictors. Note: $R^2 = 0.49$ for step 1 and $\Delta R^2 = 0.35$ for step 2. (***: p<0.001)

	B	**standard error**	$\boldsymbol{\beta}$
1st step			
constant	-0.19	0.15	non signif.
1st dimension	-0.689	0.153	-0.7***
2nd step			
constant	0.007	0.087	non signif.
1st dimension	-0.628	0.089	-0.638***
3rd dimension	-0.561	0.086	-0.593***

Figure 6.3 presents the optimally rotated 3D space by depicting its three 2D planes. The different symbols for each sound represent classes of musical instruments according to von Hornbostel and Sachs [1914] and the filling of the symbols represents the type of excitation (black for continuant sounds and white for impulsive sounds). The number next to the instrument abbreviation indicates pitch height with 1 to 4 corresponding to A1 to A4. Sub-figure 6.3 (b) also includes the regression line from Equation 6.1 that represents the *texture* dimension. Sounds po-

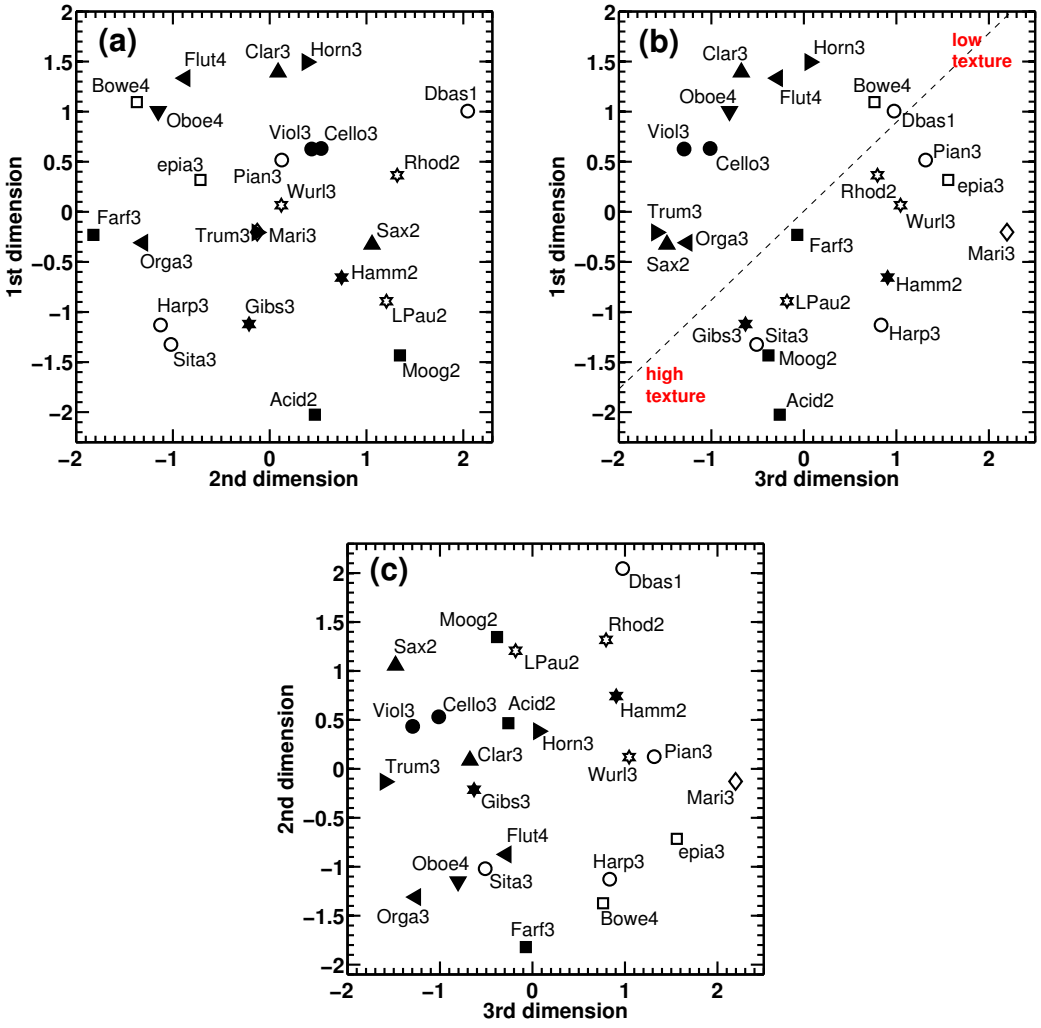


Figure 6.3: Three 2D planes of the optimally rotated 3D MDS timbre space. Black symbols: Continuant, white symbols: Impulsive, △: Single reed, ▽: Double reed, ◁: Aerophone, ▷: Lip reed, ◯: Chordophone, ◇: Idiophone, ⋆: Electrophone , □: Synthesiser. The number next to the instrument abbreviation indicates pitch height with 1 to 4 corresponding to A1 to A4. The dotted line in sub-figure (b) is the regression line of equation 6.1 which represents the auditory *texture* semantic dimension.

sitioned in the bottom left corner of this plane (e.g. Acid, sitar, Moog, Gibson guitar, saxophone, trumpet, organ etc.) are generally perceived as being rough whereas the ones in the upper right corner (e.g. double bass pizzicato, piano, Bowedpad, french horn, electric piano, marimba etc.) as being smooth. Additionally, the positioning of the sounds on the second dimension indicates both perceived *luminance* and *mass*. Sounds on the positive end of the second dimension (e.g. double bass pizzicato, Rhodes piano, saxophone, Les Paul Gibson guitar, Moog, etc.) are gener-

ally perceived as dull and thick while sounds on the negative end (e.g. Farfisa, Bowedpad, organ, oboe, harpsichord, sitar, etc.) as bright and thin.

Furthermore, Figure 6.3 shows that same-family instrument sounds cluster together in many cases. For example, the wind instruments in sub-figure 6.3 (b) form two clusters: clarinet, oboe, flute, French horn and trumpet, organ, saxophone. The cello and the violin, the only continuant chordophones, are very closely grouped in all planes. Finally, dimensions 2 and 3 seem to be affected by pitch and impulsiveness respectively as will be further supported by the next section on acoustic correlates.

6.4 Acoustic correlates of perceptual dimensions

Similarly to chapter 5, a large set of acoustic descriptors was extracted from the stimuli in order to identify acoustic correlates for the perceptual dimensions obtained by MDS analysis. High multicollinearity within our acoustic features set was also addressed as described in section 5.5. The final solution consisted of 4 components (KMO = 0.642, Bartlett's test of sphericity $p < 0.001$) that explained 83.2% of the total variance. Table 6.4 shows the loadings of the features on the 4 components after orthogonal Varimax rotation. The components' labelling is based on the acoustic correlates that are highly loaded on each one. For an explanation of the features abbreviations see Table 5.8.

Features like the *normalised harmonic spectral centroid* (SC_norm), *tristimulus 3* (T3) [Pollard and Jansson, 1982], *SC_loud_cor* (corrected version of the spectral centroid calculated from Moore's specific loudness in order to remove the influence of F_0, for an example see Marozeau and de Cheveignè [2007]) all represent spectral structure (i.e. distribution of energy among harmonic partials) rather than spectral content. Therefore, the first component is labelled: *energy distribution of harmonic partials*. The second component is represented by both *odd even ratio* (OER) and *inharmonicity*. The third component is related to *spectrotemporal* characteristics such as *noisiness*, *harmonic spectral flux* (Flux) and the *standard deviation of the harmonic spectral centroid* (SC_std). Finally, the fourth component is related to temporal characteristics such as *the logarithm of the attack time* (Log_At_time) and the temporal centroid (TC) and spectrotemporal ones such as the temporal variation of the first nine harmonics (*Mean coefficient of variation, MCV,* Kendall and Carterette, 1993b)

Table 6.5 presents the Spearman correlation coefficients between the three perceptual di-

Table 6.4: Component loadings of the acoustic features on the first 4 principal components as a result of PCA with Varimax rotation. Loadings ≥ 0.7 are presented in bold and used for component labelling.

	Component			
	1	**2**	**3**	**4**
	(Energy distribution of harm. partials)	(Inharmonicity, OER)	(Spectrotemporal)	(Temporal, Spectrotemporal)
SC_loud	**0.701**	0.653	0.014	0.066
T3	**0.957**	0.060	0.-0.024	0.045
SC_loud_cor	**0.845**	0.506	0.029	0.008
SC_norm	**0.940**	0.042	0.053	-0.045
Spread	**0.730**	0.450	-0.009	-0.020
T2	**-0.931**	0.164	0.090	0.137
Inharmonicity	0.150	**-0.711**	0.426	-0.349
OER	-0.166	**-0.773**	-0.261	-0.148
Noisiness	0.238	0.083	**0.875**	-0.153
Flux	-0.055	-0.140	**0.823**	0.039
SC_std	-0.296	0.246	**0.720**	0.220
SC_var_loud	-0.625	-0.613	-0.074	-0.148
Log_At_time	0.077	-0.039	0.228	**0.880**
MCV	-0.223	-0.445	-0.016	**0.761**
TC	0.237	-0.474	-0.133	**0.744**

mensions, the four principal components of the acoustic features together with the fundamental frequency (F_0) and temporal centroid. Although temporal centroid was also loaded on the fourth component, its correlation with the third dimension is also separately reported as it demonstrates their relationship more emphatically. The energy distribution of harmonic partials seems to influence both dimensions 1 and 3 equally [$\rho(22) = -0.668$ and $\rho(22) = -0.704$ respectively, $p < 0.001$]. The second dimension correlates well with the second principal component [$\rho(22) = -0.668$, $p < 0.001$] and is additionally strongly correlated with F_0 [$\rho(22) = -0.818$, $p < 0.001$]. No significant correlation was found between any of the dimensions and the third (spectrotemporal) component and only a mild correlation was identified between the fourth component (temporal and spectrotemporal) with the first and third dimensions [$\rho(22) = 0.523$ and $\rho(22) = -0.578$ respectively, $p < 0.01$]. However, the temporal centroid alone is strongly correlated with the third dimension [$\rho(22) = -0.762$, $p < 0.001$].

Table 6.5: Spearman correlation coefficients between perceptual dimensions, the 4 principal components of the acoustic feature set plus F_0 and temporal centroid. (*: $p<0.05$), **: $p<0.01$, ***: $p<0.001$)

	Relative energy of the harmonic partials	OER Inharmonicity	Spectrotemporal	Temporal, Spectrotemporal	F_0	TC
1st Dimension	-0.668***	0.016	-0.146	0.523**	0.458*	0.313
2nd Dimension	0.069	-0.666***	0.009	0.066	-0.818***	-0.198
3rd Dimension	-0.704***	-0.279	0.064	-0.578**	-0.117	-0.762***

6.5 Discussion

In contrast to the semantic spaces of chapter 5 where there did not seem to be a clustering of sounds based on instrument family or means of excitation, same-family instruments occupied similar regions in this perceptual space. This is in agreement with the literature of pairwise dissimilarity experiments [e.g. Giordano and McAdams, 2010]. Additionally, F_0 and temporal centroid were strongly correlated with dimensions 2 and 3 respectively.

However, this study also provides evidence that verbal description and pairwise comparison can result in related representations of musical timbre based on the correlation analysis between semantic and perceptual dimensions. The fit between semantic and perceptual spaces was improved compared to previous studies [e.g. Kendall and Carterette, 1993a,b, Kendall et al., 1999], a fact that could be attributed to the analytic treatment of verbal descriptions. More specifically, nonlinear relationships between semantic variables were accounted for through optimal variable transformations and a more easily interpretable non-orthogonal rotation was applied to the semantic dimensions identified by factor analysis (see chapter 5). Auditory *luminance* featured a strong correlation with the second MDS dimension. Also, auditory *texture* was significantly correlated with two of the MDS space dimensions (first and third). A stepwise multiple regression attributed almost equal importance to each of the two dimensions in determining position on *texture* dimension. Auditory *mass* showed strong correlation with the same dimension as auditory *luminance* (second). This implies that the MDS perceptual timbre space was not able to account for the unique variance of either *luminance* or *mass*. These high correlations were found despite the differences in durations (shorter in pairwise dissimilarity) and slight alteration of some pitches between the stimuli of the two experiments. The stability of instrument perception regardless of duration has also been noted by Kendall et al. [1999].

It seemed possible for participants to make judgements of timbral dissimilarity even for an F_0 range of three octaves, but at the same time F_0 variation explained more than 65% of the variance on one of the MDS dimensions, supporting Marozeau et al. [2003] and Marozeau and de Cheveignè [2007]. However, F_0 variation was by no means overshadowing every other timbral dimension as has been reported for simple synthetic stimuli [Miller and Carterette, 1975]. It could be argued that the timbral complexity of natural sounds prevailed over a wide range of F_0s. Furthermore, F_0 seemed to significantly influence the perceived *mass* and *luminance*, confirming the findings of chapter 5. F_0 positively contributed to *luminance* perception which also supports [Marozeau and de Cheveignè, 2007] and [Schubert and Wolfe, 2006]. However, a corrected calculation of *SC* according to Marozeau and de Cheveignè [2007] did not confirm that it could be a better predictor of auditory *luminance*. Previous indications (see chapter 5) that *inharmonicity* is an acoustic correlate for auditory *mass* and *luminance* and also that the energy distribution of harmonic partials is a good predictor for auditory *texture* were supported. Finally, it has to be noted that the third MDS dimension seemed to additionally differentiate between percussive and continuant instruments as indicated by the strong correlation with the temporal centroid.

6.6 Conclusion

The purpose of this chapter was to evaluate semantic description of musical timbre. To this end, a semantic timbre space that resulted from a verbal magnitude estimation listening test was compared with a perceptual timbre space that came from a pairwise dissimilarity rating listening test. Both these timbre spaces concerned the same sound stimuli. The comparison revealed a considerable degree of fit between the projections on perceptual and semantic dimensions. This finding supports the idea that the three salient semantic dimensions (*luminance*, *texture*, *mass*) can, to some extent, capture the perceptual structure of a set of timbres, thus implying a critical latent influence of timbre semantics on pairwise dissimilarity judgements. In other words, the perceived dissimilarity between a pair of different timbres might be influenced by the integration of a number of subconscious evaluations on several latent semantic dimensions. Further research is required, however, to examine the level of independence between *luminance* and *mass*. Finally, the correlation of the energy distribution of harmonic partials with auditory *texture* and the association of inharmonicity and F_0 to auditory *luminance* and *mass* was further supported.

Chapter 7

Partial timbre

7.1 Introduction

In the previous two chapters we have not only shown that the salient semantic dimensions of timbre feature strong similarities between English and Greek but also that they convey a substantial amount of perceptual information. In other words, certain perceptual attributes of musical timbre can be reflected through verbal description. Since we have found that description of musical timbre is meaningful, we now examine whether timbral relationships among sounds are affected by the auditory environment.

Everyday experience shows that even the listening level affects the way we perceive music. When playback level is increased (within a reasonable range), then frequencies that were previously inaudible come into play. This phenomenon usually affects lower and higher frequencies more as a result of our lower sensitivity for this part of the spectrum. For example, when someone listens to a Mahler symphony[1] at a comfortable level, it is possible that he or she misses some amount of timbral richness (or even melodic and rhythmic information), especially at parts of lower orchestra dynamics (i.e. quieter passages or even at louder passages performed by low register instruments such as the cellos). Therefore, our perception of a musical piece is dependent on the listening level to such an extent that it may even affect the composer's original intentions.

An analogous effect can be produced by auditory masking. In a complex auditory environment the concurrent presence of several sound sources can result in some of them becoming

[1] Mahler symphonies are used as an example for their wide dynamic range and timbral richness.

barely audible or completely inaudible [Fastl and Zwicker, 2007, Moore, 2003]. Real life auditory scenes can consist of both competing sound sources and background interference. Sounds (or their portions) that are below the masking threshold are usually severely affected. Listening to music in noisy environments is quite common. The noise of the engine and the tyre friction when listening to music in a car, the background noise when using headphones outdoors, even the noise coming from an open window when listening to music in our living room are only but a few examples of background interference. Likewise, sounds that are constructively combined also interact with each other. That is, masking can also take place within a musical ensemble (i.e., the presence of a dominant audio stream may mask parts or the entirety of other concurrent audio streams). A conductor in live music performance or a mixing engineer in recorded music can control, among various other things, the relative levels among instruments in order to achieve the desired sonic result. In general, the masking mechanism is the same either for background interference or for constructive combination of sound sources.

Moore et al. [1997] has used the term *partial loudness* to refer to the contribution of a single sound source to the overall loudness of a mix of concurrent sounds. The existence of partial loudness implies a certain degree of distinctness of a sound in a mixture (if a sound cannot be even slightly distinguished from the background then its partial loudness is eliminated). The concept of partiality could be extended to timbre, where *partial timbre* would refer to the portion of the original timbre (i.e. timbre in isolation) that is retained in a sound when heard in the presence of other sounds[2]. For example, a guitar would probably feature a different timbre as part of a densely textured rock ensemble than if heard in isolation. As mentioned above, this will be due to masking caused by the other competing sounds. Proportionally, the timbral semantics of a sound in isolation might differ compared to it being heard as part of a complex auditory scene, implying that timbral qualities are context dependent rather than absolute.

As a first step to test the hypothesis that timbre perception is significantly affected by the auditory environment, we focused on the perceived timbral differences as a result of interfering background noise. White noise was the favoured masker since, despite not being musical, it represents a general, broadband and easily reproducible masker that can clearly demonstrate the existence of a potential effect. A pairwise dissimilarity rating listening test with three different listening conditions was designed and conducted. It involved the pairwise comparison of 13

[2]The term *partial* in this case is used as an adjective referring to the part of timbre and should not be confused with a harmonic partial.

synthesised sounds in silence and under two different levels of background white noise. The data were analysed through non-metric MDS (see subsection 3.1.1) and the resulting timbre spaces were subsequently compared using cluster analysis (see subsection 3.1.2).

7.2 Method

7.2.1 Stimuli and apparatus

Thirteen complex, tonal sounds were synthesised using a custom made additive synthesiser in Max/MSP. The synthesiser offered thirty nominal partials, which could be independently controlled for: maximum amplitude, Attack-Decay-Sustain-Release (ADSR) type envelope, amplitude and frequency modulation, inharmonic displacement and phase[3]. Figure 7.1 shows the partial level diagram of the synthesiser. The synthesiser parameters were exploited to create stimuli with the characteristics of real-world musical sounds (i.e. having various spectral profiles, temporal envelopes, spectrotemporal variations and inharmonicities). Each sound was 600 ms long and the inter stimuli interval was 400 ms. F_0 was kept constant at 392 Hz (G4). The spectrograms of the 13 sound stimuli in *silence* condition shown in Figure 7.2, demonstrate the significant timbral variability within the sound set.

Prior to the listening test, the stimuli were equalised in loudness. Within each condition (*silence*, *low-noise*, *high-noise*), the stimuli were each adjusted repeatedly in level through an informal listening test within the research team until equal loudness was achieved across all stimuli. I.e., the levels were adjusted separately for each condition. In each condition containing background noise, real-time generated white noise was presented continuously throughout the block. Figure 7.3 indicatively shows the effect of background noise on the spectrograms of sound stimuli No. 5, 9 and 12.

The levels of the target sounds (i.e., not including the background noise) and the background noise were selected so as to provide a comfortable listening experience for the silence condition and two distinct background masking conditions. In all three conditions, the listening level of the target sounds was measured to be approximately 60 dB SPL (RMS). The background noise level was measured at 44 and 68 dB SPL (RMS) for the *low-noise* and *high-noise* conditions respectively.

In a post-test questionnaire, all the participants reported that the level was comfortable for

[3]Phase alterations were not utilised for the purposes of this experiment.

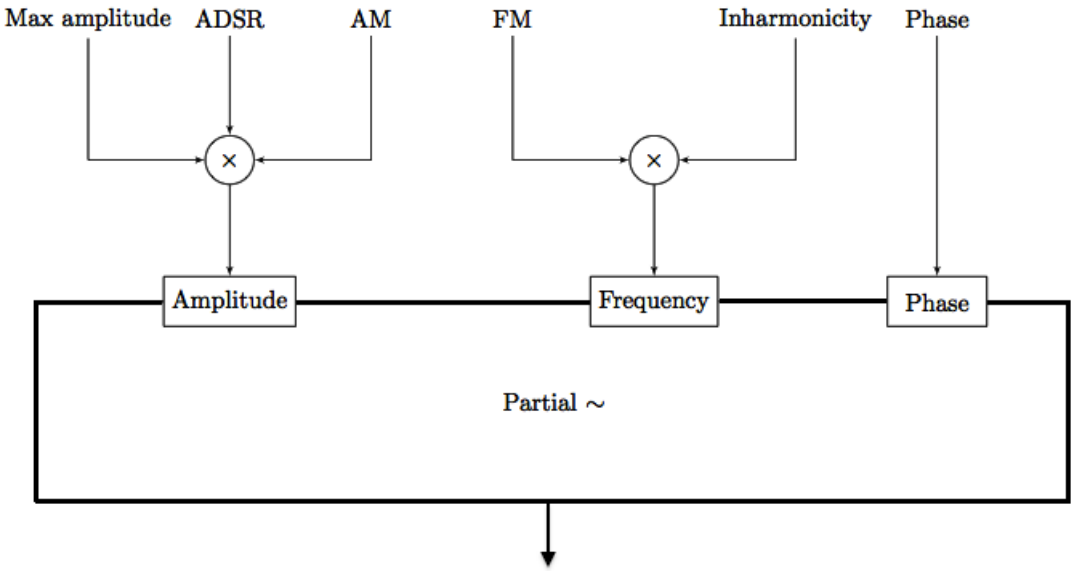


Figure 7.1: Partial level diagram of the additive synthesiser. The amplitude of each partial is defined by a combination of maximum amplitude, ADSR envelope and sinusoidal amplitude modulation. The exact frequency position of each partial is defined by an initial displacement of the harmonic position together with a sinusoidal frequency modulation. Phase takes an angle from 0° to 360° as an input.

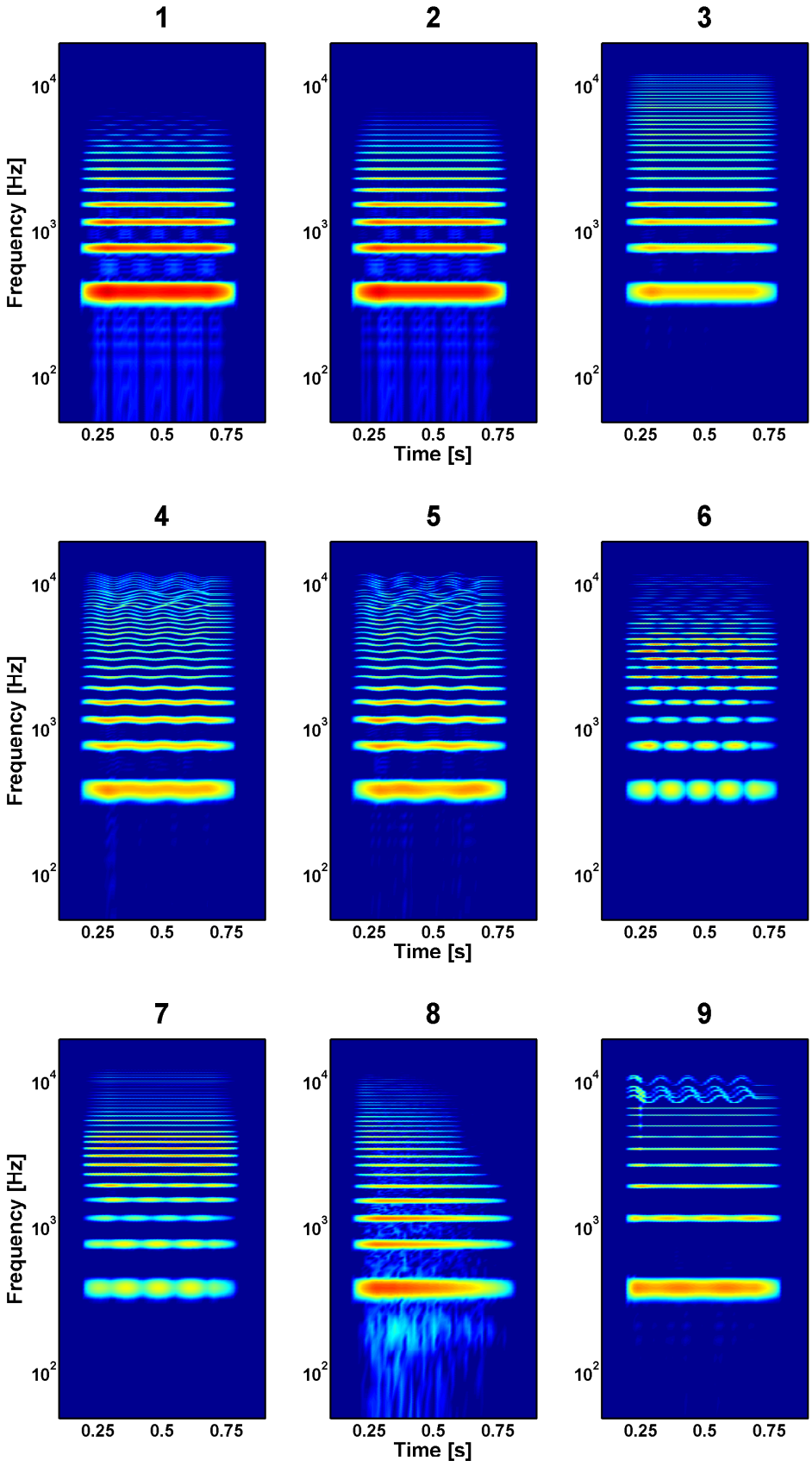
all stimuli and confirmed that loudness was constant within blocks (i.e., within conditions) and across stimuli. They also reported that the target sounds were somewhat quieter in *low-noise* and considerably quieter in *high-noise* conditions (though never inaudible). The listening test was conducted under controlled conditions in an acoustically isolated listening room. Sound stimuli were presented through the use of a laptop computer, with a Tascam US122L external audio interface and a pair of Sennheiser HD60 ovation circumaural headphones.

7.2.2 Participants

Nine volunteer participants (aged 22-41, mean age 29, 3 female) participated in the listening test. All reported normal hearing and long term music practice (17.2 years on average, range: 10 to 25). Participants were researchers from the Centre for Digital Music at Queen Mary University of London. All participants were naive about the purpose of the test.

7.2.3 Procedure

Paired sounds were presented in blocks of 91 trials. Each listener completed one block in each of the three conditions; *silence*, *low-noise* and *high-noise*. Blocks were presented in random order. Trials within blocks were selected in random order, and presentation order of the paired sounds


1
2
3
4
5
6
7
8
9
Frequency [Hz]
10^4
10^3
10^2
0.25
0.5
0.75
Time [s]

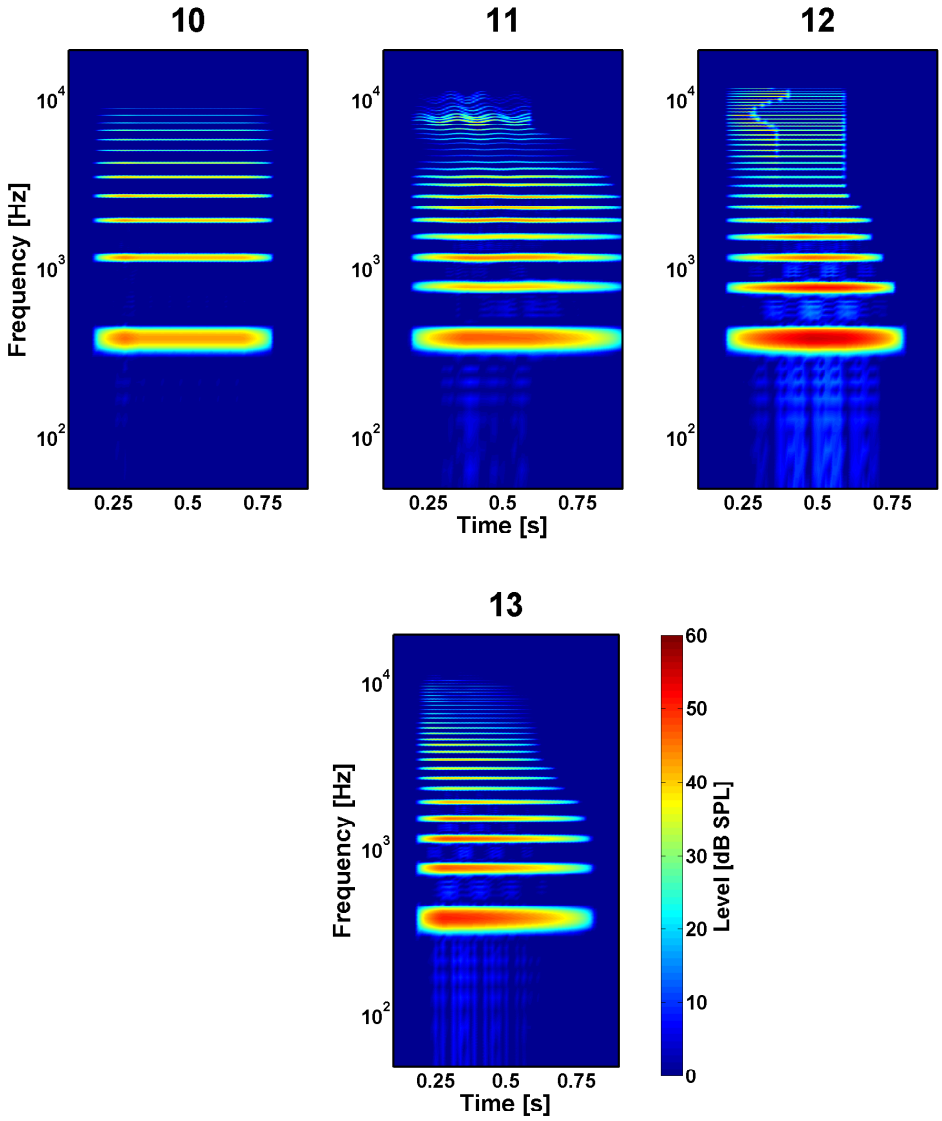
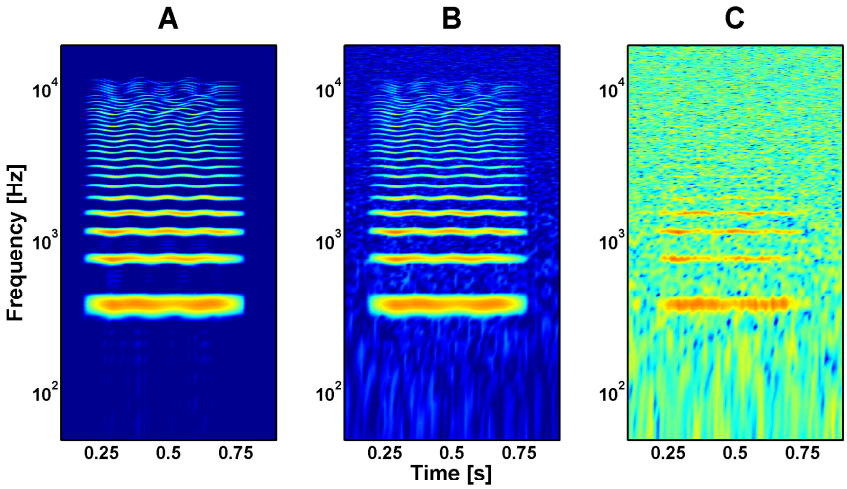


Figure 7.2: Stimuli spectrograms illustrating the spectrotemporal features of the stimuli. Panels **1** - **13** show the spectrograms of the thirteen respective sounds in the *silence* condition.

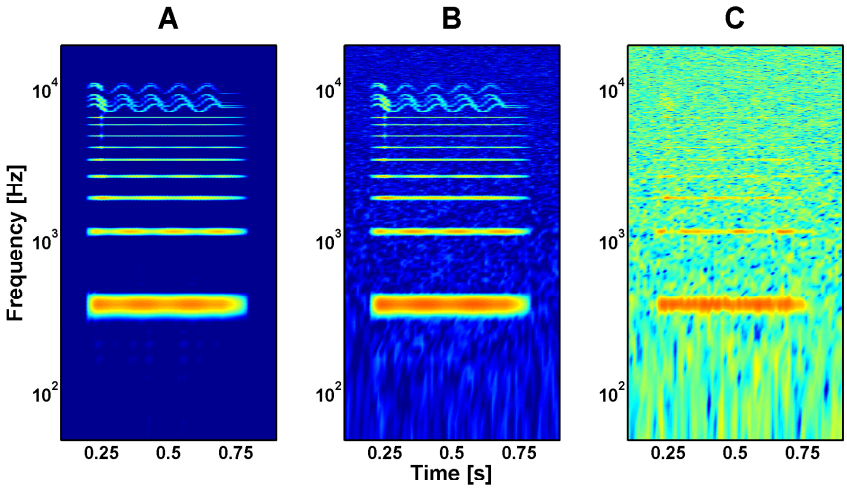
was also randomised. All pairwise combinations of the thirteen sounds were presented, including *same-sound* pairs.

Prior to each block, each listener was presented with the entire set of stimuli (within that condition) at random, in order to become familiar with the overall dissimilarity range. This was followed by a brief training session where listeners completed part of a block. The training data were discarded. Similarly to what was described in chapter 6, listeners rated the perceptual distances between pairs by freely inserting a number of their choice for each pair with 0 indicating an identical pair. The ratings were then normalised for each listener. Listeners were advised to maintain a consistent rating strategy throughout the experiment.

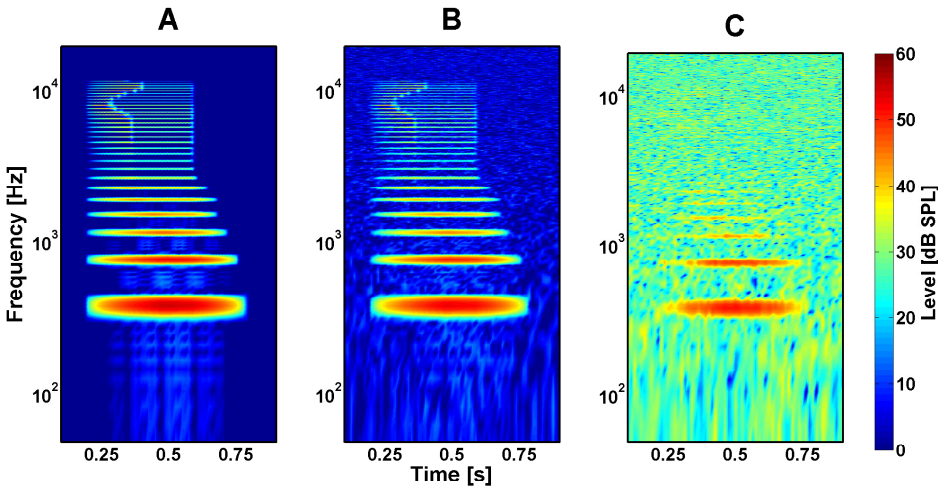
For each trial, listeners were permitted to listen to each pair of sounds as many times as



(a) Sound stimulus No. 5.



(b) Sound stimulus No. 9.



(c) Sound stimulus No. 12.

Figure 7.3: Background noise spectrograms showing the effect of background noise on typical stimuli (sound indices 5, 9 and 12 are represented by sub-figures (a), (b) and (c) correspondingly). **A** shows the spectrogram of the sound in the *silence* condition. **B** shows the spectrogram of the sound in the *low-noise* condition. **C** shows the spectrogram of the sound in the *high-noise* condition.

necessary before submitting their dissimilarity rating. Listeners were also encouraged to take regular breaks and were free to do so at any time. The overall listening test procedure, including instructions, lasted around one hour and a half for most of the participants.

7.3 Results

The Cronbach's Alpha (see subsection 3.1.5) among participants exceeded 0.8 for the *silence* and *low-noise* conditions and 0.9 for the *high-noise* condition indicating high inter-participant reliability. Multidimensional scaling (the weighted Euclidean PROXSCAL algorithm presented in subsection 3.1.1) has been utilised to construct the geometric configuration of our stimuli timbre space, which allowed interpretation of dissimilarity data by Euclidean methods, e.g. the correlation between the spaces and differences in their structure. Non-metric MDS analysis [Kruskal, 1964a,b, Shepard, 1966, Young, 1970] was initially performed over a range of dimensionalities to determine the order most suitable to represent the timbre space for each presentation condition. Table 7.1 shows the evolution of two measures-of-fit (*S-Stress* and *DAF*) of the PROXSCAL algorithm for orders of dimensionality between one and four. The improvement of the measures-of-fit from a 3D to a 4D space was minimal and hence three dimensions were deemed optimal to represent the data for all background conditions.

Table 7.1: Measures-of-fit for different MDS dimensionalities for *silence*, *low-noise* and *high-noise* conditions.

Condition	Dimensionality	**S-Stress**	Improv.	**DAF**	Improv.
silence	1D	0.357	–	0.825	–
	2D	0.167	0.190	0.9133	0.0883
	3D	0.092	0.075	0.968	0.0547
	4D	0.055	0.037	0.983	0.015
low-noise	1D	0.371	–	0.835	–
	2D	0.184	0.187	0.935	0.100
	3D	0.098	0.086	0.968	0.033
	4D	0.060	0.038	0.981	0.013
high-noise	1D	0.180	–	0.902	–
	2D	0.150	0.030	0.934	0.032
	3D	0.063	0.087	0.977	0.043
	4D	0.040	0.023	0.985	0.008

7.3.1 Timbre space correlations

Euclidean pairwise distances for all sounds were calculated from the timbre spaces and were correlated (Spearman) across conditions (see Table 7.2). It is evident that while the *low-noise* timbre space is relatively close to the *silence* space [$\rho = 0.79$, $p < 0.001$], the *high-noise* space shows only a mediocre correlation [$\rho = 0.54$, $p < 0.001$] with the *silence* space. This means that while the *silence* and *low-noise* spaces have 62% of their variance in common, the *silence* and *high-noise* spaces share only 28% of their variance. The *high-noise* and *low-noise* spaces are also moderately correlated [$\rho = 0.53$, $p < 0.001$]. In other words, the timbre spaces in the respective conditions are fundamentally different and the timbre space resulting from the *high-noise* condition is most different to that of the *silence* condition.

Table 7.2: Spearman correlation coefficients of pairwise distances between the timbre spaces for the three different conditions. *** : $p < 0.001$

	silence	low-noise	high-noise
silence	1.0	-	-
low-noise	0.79^{***}	1.0	-
high-noise	0.54^{***}	0.53^{***}	1.0

7.3.2 Structural changes in timbre spaces

An average linkage hierarchical cluster analysis (see subsection 3.1.2) was performed on the 3D coordinates of the timbre spaces and yielded the dendrograms shown in Figure 7.4. We applied the method given in Morlini and Zani [2012] to determine the similarity in the structure of the timbre spaces across presentation condition. This method constructs the matrix X of binary values which describes the grouping of each stimulus pair for all non-trivial[4] numbers of clusters, e.g. $X_{i,j}$ shows whether the stimuli in pair i are in the same cluster when j clusters were considered; and with 13 stimuli our analysis was performed with 2-12 clusters. The dissimilarity index Z is calculated by comparing X matrices derived from two sets of data, and we apply this to the timbre spaces for each pair of presentation conditions. The similarity scores (evaluated by taking the compliment of 1 to Z) are shown in Table 7.3, where a value of 1 represents identical structure and a value of 0 identifies the maximum degree of dissimilarity. It shows that there are

[4]If the number of clusters is 1, or is equal to the number of stimuli, the solution is considered to be trivial because all stimuli will be in either the same or individual clusters, and hence no differences between spaces can exist.

differences in the structural grouping of stimuli across the presentation conditions, and that the differences become greater as the level of background noise is increased from silence.

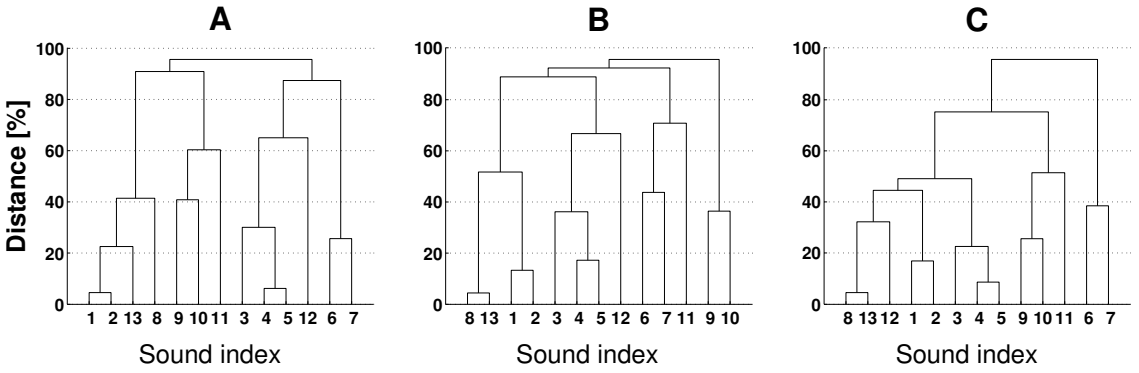


Figure 7.4: Dendrograms from hierarchical cluster analysis of *silence* (**A**), *low-noise* (**B**) and *high-noise* (**C**) conditions. The index numbers on the abscissa represent the thirteen stimuli used for the experiment.

Table 7.3: The structural similarity in the timbre spaces across the three background conditions.

	silence	low-noise	high-noise
silence	1.0	-	-
low-noise	0.62	1.0	-
high-noise	0.60	0.71	1.0

7.4 Discussion

The main goal of the study presented in this chapter was to provide some insight into timbre perception in more realistic scenarios such as background noise interference or simultaneous presence of multiple sound sources, e.g. music. As a first step, we compared dissimilarity ratings among synthetic sounds in different levels of background noise. We have shown that the presentation condition caused significant changes to the timbre space. Interestingly, the presence of noise altered structural relationships within the timbre spaces rather than just causing a simple contraction or expansion. This means that stimuli grouped in one presentation condition may be perceived as being unrelated in another and vice versa.

We suggest that changes in the timbre spaces can be attributed to either background noise features being incorporated into (fused with) the target sound, or to features of the target sound being incorporated into the background noise. The most striking changes in the cluster structure

(Fig. 7.4) between conditions were for sound pair 8-13 with the introduction of *low* or *high-noise*; sound pairs 9-10 and 6-7 and for sound 12 with the introduction of *high-noise*. For example, sound 8 differs from sound 13 only by an added white noise component (see Fig. 7.2) which seems to be grouped with background noise in *low* or *high-noise* conditions, thus making 8 and 13 indistinguishable. The tighter clustering of sounds 9-10 and the clustering of sound 12 with pair 8-13 indicates that the frequency modulation and inharmonicity present in the higher partials of sound 9 as well as the characteristic specificity (shorter attack time and faster decay of the higher partials) of sound 12, were all obscured in the *high-noise* condition. This suggests that the randomness and inharmonicity inherent in the background noise present an ambiguity to the listener, since he or she is unable to determine whether these auditory features are attributable to the background noise or the target sound. Finally, the *high-noise* dendrogram of Figure 7.4(C) demonstrates that while sound pair 6-7 is evidently an outlier, the rest of the stimuli cluster together more tightly (i.e. are less distinguishable from each other). It seems that the higher concentration of energy between partials 5 and 12 (1.97kHz-4.74kHz) together with the strong amplitude modulation featured in both 6 and 7 have largely allowed them to retain their distinct identity relative to the rest of the stimuli.

Whilst the synthetic sounds we employed had the characteristics of real-world musical sounds, they did not resemble specific instruments, so they are not likely to be subject to higher level and/or more abstract categorical cues to similarity, e.g. “this sound is a piano”. Hence it seems unlikely that high level (abstract) informational masking might have played a role.

7.5 Conclusion

The above findings confirm that the timbre of a sound is not an absolute percept but is instead related to the auditory environment in which it is being experienced. Based on this we propose the use of the term *partial timbre* to describe the portion of the original timbre (i.e. timbre in isolation) that is retained when a sound is heard in the presence of other sounds. This definition of *partial timbre* presupposes *timbral heterogeneity* of the sound sources and examines the influence of an auditory scene on the timbre of a specific sound source. As a future work and in direct analogy to *partial loudness*, it would be interesting to also examine the way in which each timbral component of an *emerging timbre* (see 2.7.2) contributes to the overall timbre. This would require a number of stimuli whose timbres are perceptually fused into one single entity.

Both music creation and music ‘consumption’ could benefit from the modelling of timbral interactions. Potential applications could range from music composition and production to areas related to sound reproduction in real-world environments.

Chapter 8

Conclusion and further work

8.1 Relationship of perception with semantics

This thesis addressed several aspects of musical timbre perception and its semantic description. The main aim of this work was to explore verbal communication regarding the quality of individual musical sounds and define a semantic framework for timbre description.

Chapter 4 described an initial approach on timbre semantics according to which conclusions would be reached through direct manipulation of synthesised stimuli parameters. The assumption that modifying specific acoustic characteristics attributed by the literature to auditory brightness and warmth would correspondingly influence their perception could not be clearly validated. The inconclusive results of this experiment indicated that the investigation of timbral semantics would be better served by a holistic rather than a narrowly focused approach. This experience led to the design of a number of experiments including musical tones to explore the perception and semantics of timbre. Targeting at the formulation of a semantic framework for timbre description, these experiments were designed so as to address a number of questions.

The first question concerned the meaning of semantic description; i.e. whether verbal description can reflect what is actually perceived in a consistent manner. If it turned out that description did not match perception at all then the investigation of timbral semantics would be pointless. The second question was about universality of timbral semantics. Supposing that timbral semantics are significantly conditioned by language of description then a separate semantic framework should be defined for each different language.

Chapter 5 described an experiment that addressed the question of semantic universality. Two groups of native Greek and English speakers described (using their native language) a set of timbres along 30 predefined semantic scales. The reduction of these high dimensional data by means of factor analysis was based on an approach that differed from the usual practice in two ways. The first was the optimal nonlinear transformation of the semantic variables which represented them in a more compact manner (in terms of cluster analysis) and which also accounted for greater percentage of total factorial variance. This approach demonstrated that semantic variables for timbre description are not guaranteed to be linearly related and they should not be treated as such. The second way was the fact that the resulting factors (i.e. semantic dimensions) were allowed to be correlated by employing a final non-orthogonal rotation. By minimising such restrictions we have enhanced the interpretability of the final solution.

Three salient semantic dimensions that accounted for over 80% of the variance were identified for each linguistic group. The conceptually related dimensions for both languages not only featured significant correlations but two-sample Kolmogorov-Smirnov tests also showed no effect of language. At the same time, some degree of ambiguity was introduced by the fact that the conceptually related semantic dimensions did not merely feature a straightforward one to one mapping. Some more complex relationships between them could potentially be attributed to a mild effect of language of description which should be further investigated. However, taken as a whole these findings supported the hypothesis of universality regarding timbral semantics, at least based on the evidence from two European languages, and demonstrated that semantic spaces exhibit three salient dimensions which we have labelled as *luminance*, *texture* and *mass*.

The encouraging findings regarding universality of timbral semantics paved the way for the second major experiment of this work. This experiment investigated the amount of perceptual information conveyed by timbral description. As discussed in chapter 6, there already existed some evidence, coming from a variety of disciplines (i.e. psychoacoustics, linguistics, neuroscience), that verbal description is perceptually meaningful. Still, all the previous attempts to link a semantic to a perceptual space through a psychoacoustics approach had revealed only partial similarities [Kendall and Carterette, 1993a,b, Kendall et al., 1999].

As also mentioned in chapter 2, the perceptual spaces[1] resulting from pairwise dissimilarity tests and subsequent MDS analyses are usually characterised by same-family instrument clusters.

[1] Such spaces are characterised as perceptual rather than semantic since the process through which they are obtained does not include any form of lexical description.

That is, their spatial structure basically categorises sounds according to their sources. This was, indeed, the case for our perceptual space as well. However, our semantic spaces did not demonstrate such an organisation. This should be expected since the adjectives used for description were focused on *iconic musical meaning* and *sound impression* rather than source description. Additionally, it is reasonable to assume that lexical description is not adequate for describing every perceivable aspect of sound.

Chapter 6 described the comparison between the English semantic space and a perceptual space obtained from different participants on the same stimulus set. Both semantic and perceptual spaces were 3-dimensional and the similarities between the semantic and perceptual dimensions were strong. One of the perceptual dimensions was found to be highly correlated with both *luminance* and *mass*, and position on *texture* could be equally determined by the other two perceptual dimensions as indicated by a multiple regression analysis. We argue that the increased similarity between perception and semantics that was evident in our work in comparison to previous studies could be attributed to a more systematic treatment of our semantic variables.

On the whole, considering the variety of stimuli (both continuant and impulsive, acoustic and synthesised, chordophones and winds etc.) and pitches used in the experiments, the similarities identified between the perceptual and semantic space were quite strong. Had the range of stimuli under test been limited (e.g. only continuant acoustic instruments) the relationship between the spaces might have been even stronger. This set of experiments not only demonstrated consistency of timbre lexical description across two different languages but also showed that it can convey a substantial amount of perceptual information.

8.2 Acoustic correlates of semantic dimensions

These experiments have also allowed us to identify some acoustic correlates of the semantic dimensions. We have found strong evidence that the energy distribution of harmonic partials is related to auditory *texture*, i.e. the more energy concentrated in the upper partials the harsher a sound is perceived to be and vice versa. Both auditory *luminance* and *mass* seem to be affected by F_0 and inharmonicity and there has also been some evidence that they may be associated with spectrotemporal variation.

Despite these findings, we have by no means come up with definitive conclusions regarding the physical properties of semantic dimensions. There are several reasons why this is particu-

larly challenging. First of all, the studies that form this thesis have not examined specificities of musical sounds. It may well be the case that the positioning of a sound along a semantic dimension comes as a result of a unique physical characteristic and cannot be explained in terms of physical properties shared with the rest of the stimuli in a set. A second factor that may also blur the picture is that the perception of dynamic entities, such as musical sounds, cannot be adequately represented by static audio descriptors, i.e. global features or descriptive statistics of time-varying features. This fact has also been pointed out by some of the listening test participants who, in subsequent informal discussions, informed us that they had applied two seemingly contradictory semantic descriptors for certain sound stimuli. As they explained, this was because some of the sounds developed in a semantically opposite manner compared to their beginning. Adding to the above confound, our experience shows that the same audio descriptors can vary significantly as a result of the signal representation (i.e. FFT, ERB, harmonic amplitudes) or of the various parameters of the extraction algorithm (see introductory paragraph of section 3.2). For example there were differences between descriptors calculated from the MIR Toolbox [Lartillot et al., 2008] and Timbre Toolbox [Peeters et al., 2011], especially in attack time extraction. Thus, we have eventually decided to extract our audio descriptors (most of which were harmonic) using the output of the SMS platform as input representation because it provided greater control over the relevant parameters.

Most of the above issues result from the fact that this work employed natural complex timbres rather than synthesised tones manipulated directly for the needs of one particular experiment. As explained in section 4.7, this was deemed appropriate as we pursued a wealth of semantic responses rather than judgements over a limited range of specific physical properties.

8.3 Partial timbre

A considerable level of both semantic universality and similarity between semantic and perceptual dimensions have been supported for isolated sound stimuli. However, the experimental condition of single sounds in absolute silence is extremely rare in the real world where sounds are usually heard in combination with each other. A semantic framework limited to deal just with isolated sounds would be of little use. Therefore, this work concludes by addressing one final question: is timbre an absolute percept or is it related to the sonic background?

To this end, we conducted a pairwise dissimilarity listening experiment with different back-

ground noise conditions (silence, low level of white noise and high level of white noise) while keeping participants and stimuli the same. The three resulting perceptual spaces differed among conditions (especially between silence and high-noise) indicating that timbre judgements were significantly affected by background noise. Since it has been shown that timbre perception is sensitive to the auditory environment and considering the proven relationship between perception and semantics, it can be further assumed that there will be an analogous effect on timbral semantics as well. The change in perception could be attributed to separate alterations of some of the identified semantic dimensions. For example, the perceptual change of an electric bass as a result of concurrent sounding instruments within a mix, might be attributed to its diminished auditory *mass*.

The concept of timbral partiality was introduced to describe the fact that the timbre of a sound may differ depending on whether it is heard in silence or in a complex auditory environment. We defined *partial timbre* as the portion of the initial timbre (timbre in isolation) that is retained in a sound when it is heard in the presence of other sounds.

Overall, this thesis has argued that semantic description of musical timbre is meaningful based on the evidence about semantic universality and the close relationship of semantics with perception. The salient semantic dimensions of timbre have been identified along with their acoustic correlates. Finally, it has been shown that the timbre of a sound is not an absolute percept but rather it is dependent on the auditory environment.

8.4 Future research

This thesis has prepared the ground for further fascinating research in the area of musical timbre perception. According to the most common approach, musical sounds have four separate attributes namely: duration, loudness, pitch and the multidimensional timbre. The findings of this work suggest that timbre may be further broken into at least three additional semantic attributes, i.e. *luminance*, *texture* and *mass*. Provided that some reference sounds are defined and since perceptual judgments tend to be relative in nature, a measurement scale analogous to the work by Fastl and Zwicker [2007] for auditory sharpness, roughness and sensory pleasantness could be created for our three identified semantic dimensions. Thus, musical sound could be defined by a set of unidimensional attributes as shown in Figure 8.1.

However, despite the contribution of this work, a great distance still needs to be covered

in order to achieve a comprehensive semantic framework for musical timbre description. Our experiments on semantics concerned merely isolated monophonic sound stimuli. But even under this strict condition, the three semantic dimensions could only account for slightly more than 80% of the variance and the fit with the perceptual spaces was not perfect. This implies that there may be additional semantic dimensions that can capture aspects of our perception that need to be further identified.

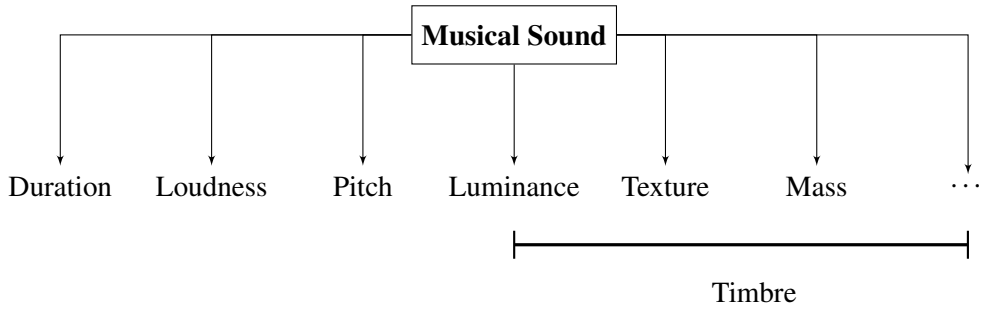


Figure 8.1: Decomposition of musical sound in its unidimensional attributes. In the case of non-pitched sounds, timbral semantics might have an even more prominent role in describing the characteristics of the sound. The dots in the final attribute imply that there might be more timbral semantic dimensions to be identified.

Furthermore, the interrelations of concurrent sounding timbres constitute a largely uncharted territory whose surface was only scratched by this work. For example, when a composer or a musician requests a particular timbre or timbral modification, it is crucial that he or she is in control regarding the effect that this new timbre will have on the entirety of the sound mix. In other words, when a particular sound quality is desired this is always in relation to the intended overall sonic outcome. Therefore, one interesting field of future research would be to investigate the contribution of each separate timbral component to the overall timbre of an auditory scene. This, of course, implies the existence of an overall timbral quality which in turn presupposes perceptual fusion. Future experiments should try to model the influence of concurrent sounds on timbre perception while demonstration of partiality regarding timbral semantic dimensions (i.e. partial *luminance*, *texture* and *mass*) could also be pursued.

All the directions described above need not be limited to a classic psychoacoustic approach. Electroencephalography, functioning magnetic resonance imaging and positron emission tomography can obtain noninvasive direct measurements of brain activity that could be proven useful on their own or in combination with traditional psychoacoustic methods.

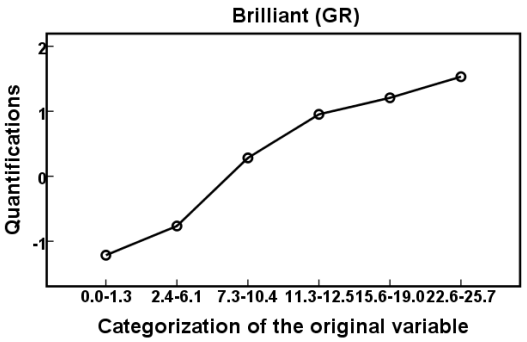
As a conclusion, research on musical timbre has the potential to result in fascinating applications that can change the way we synthesise new sounds, record, produce and reproduce music.

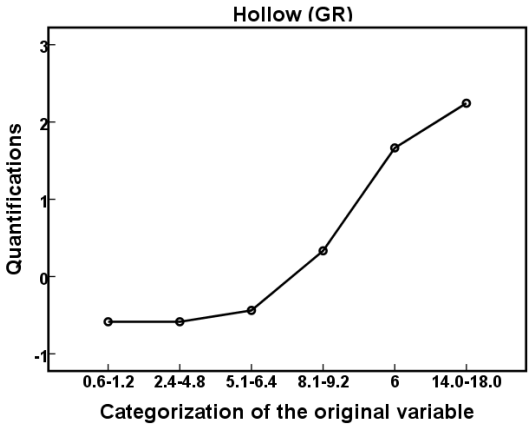
Appendix A

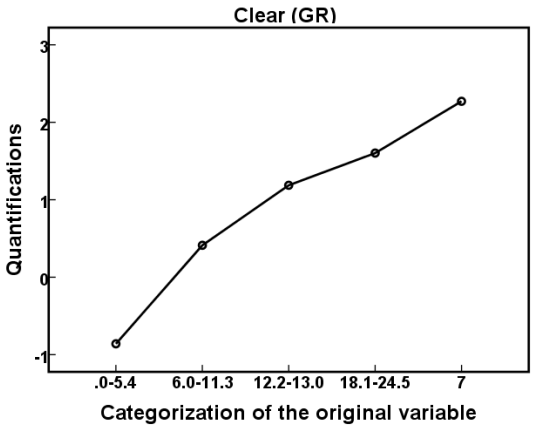
Transformation plots

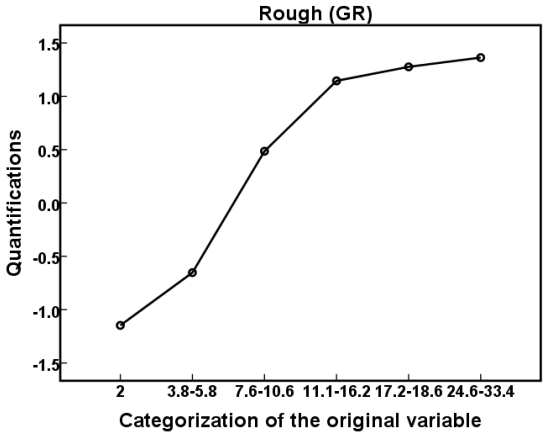
This Appendix presents the CATPCA transformation plots corresponding to each of the 30 semantic descriptors (original variables) for both Greek and English. The number of categories was initially set to 7 but sometimes the algorithm only used 6 or even 5. The optimal scaling level was set to spline ordinal (2nd degree and 2 interior knots). The *x* axes of the transformation plots show the intervals in which each variable (i.e. adjective) was categorised and the *y* axes show the value that was assigned to each category (quantification). The majority of the transformations are nonlinear, further demonstrating the usefulness of the CATPCA approach[1].

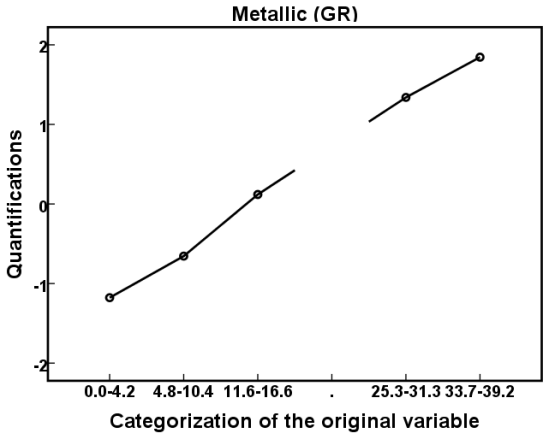
[1]Note that two indicative transformation plots are also presented in subsection 5.3.2.

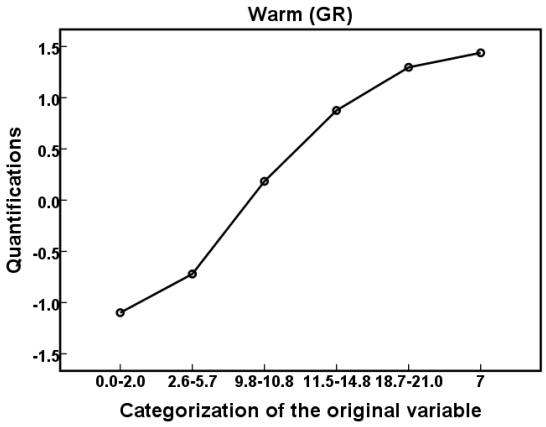

Brilliant (GR)
Quantifications
2
1
0
-1
0.0-1.3
2.4-6.1
7.3-10.4
11.3-12.5
15.6-19.0
22.6-25.7
Categorization of the original variable

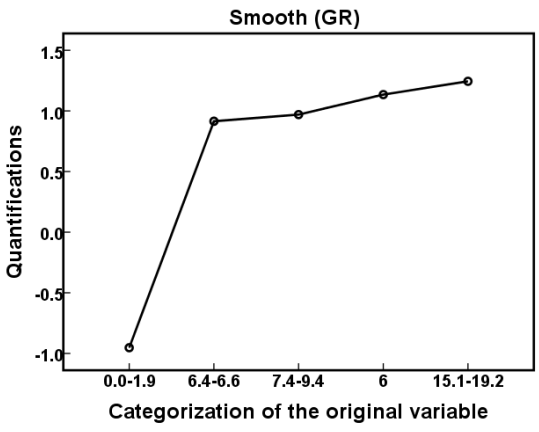

Hollow (GR)
Quantifications
3
2
1
0
-1
0.6-1.2
2.4-4.8
5.1-6.4
8.1-9.2
6
14.0-18.0
Categorization of the original variable

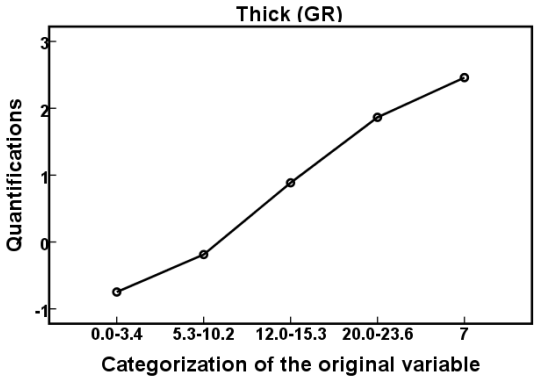

Clear (GR)
Quantifications
3
2
1
0
-1
.0-5.4
6.0-11.3
12.2-13.0
18.1-24.5
7
Categorization of the original variable

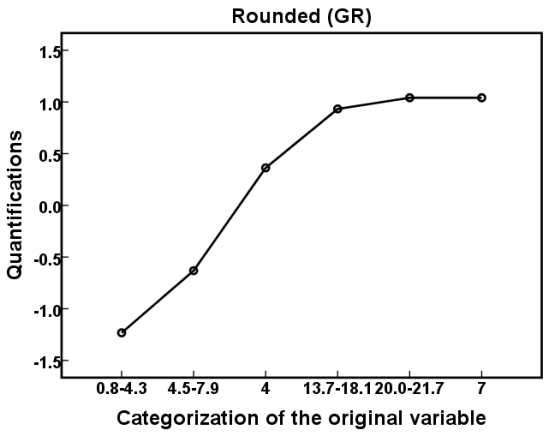

Rough (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
2
3.8-5.8
7.6-10.6
11.1-16.2
17.2-18.6
24.6-33.4
Categorization of the original variable

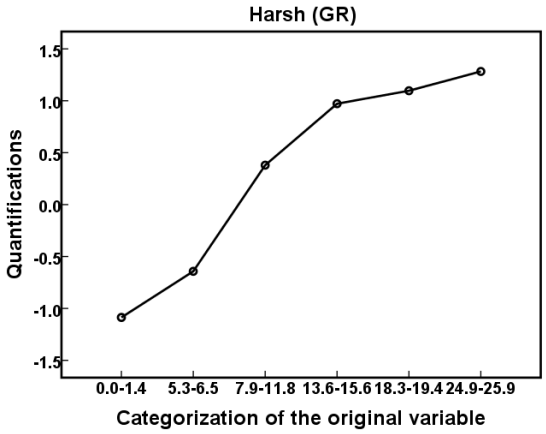

Metallic (GR)
Quantifications
2
1
0
-1
-2
0.0-4.2
4.8-10.4
11.6-16.6
.
25.3-31.3
33.7-39.2
Categorization of the original variable

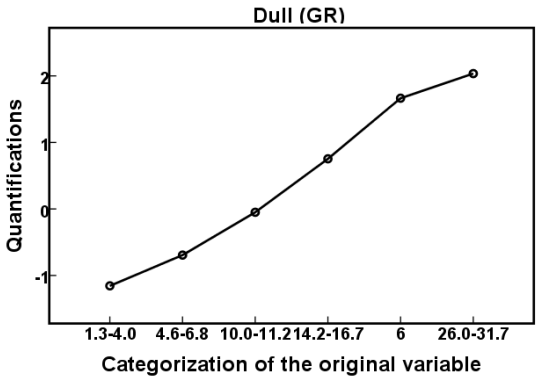

Warm (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0-2.0
2.6-5.7
9.8-10.8
11.5-14.8
18.7-21.0
7
Categorization of the original variable

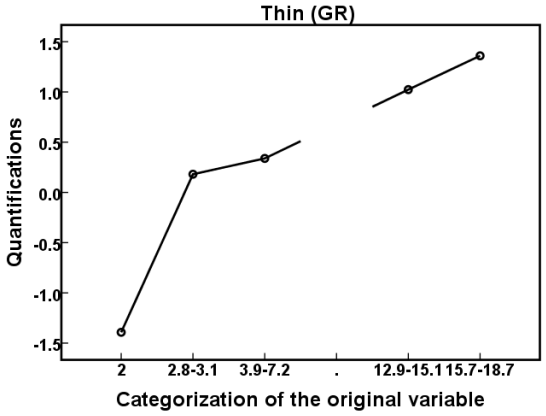

Smooth (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
0.0-1.9
6.4-6.6
7.4-9.4
6
15.1-19.2
Categorization of the original variable

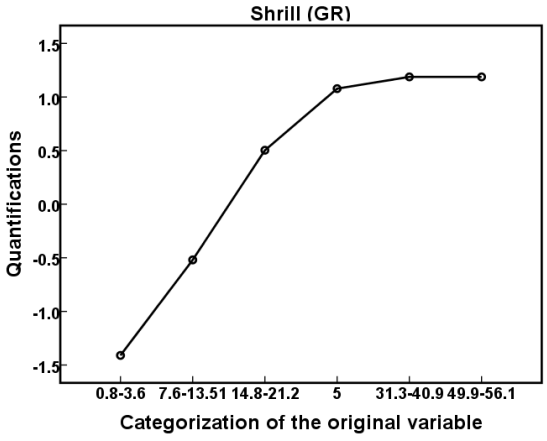

Thick (GR)
Quantifications
3
2
1
0
-1
0.0-3.4
5.3-10.2
12.0-15.3
20.0-23.6
7
Categorization of the original variable

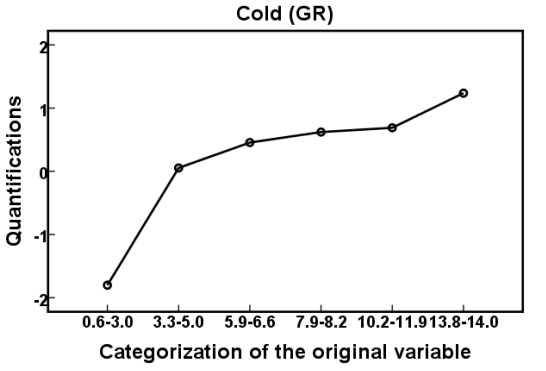

Rounded (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.8-4.3
4.5-7.9
4
13.7-18.1
20.0-21.7
7
Categorization of the original variable

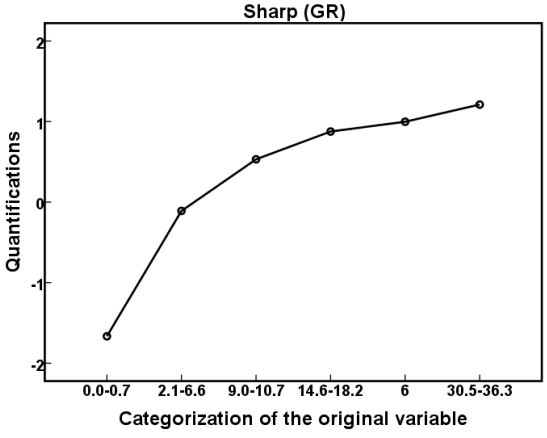

Harsh (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0-1.4
5.3-6.5
7.9-11.8
13.6-15.6
18.3-19.4
24.9-25.9
Categorization of the original variable

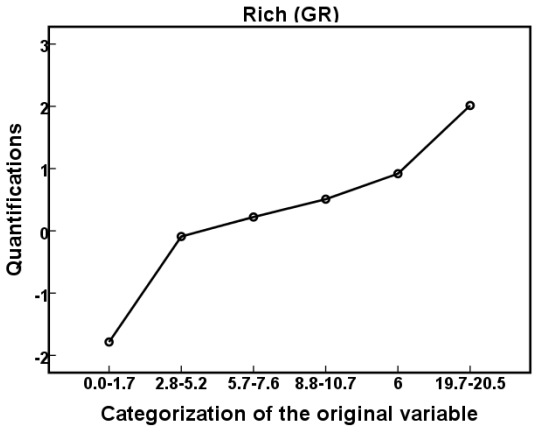

Dull (GR)
Quantifications
2
1
0
-1
1.3-4.0
4.6-6.8
10.0-11.2
14.2-16.7
6
26.0-31.7
Categorization of the original variable

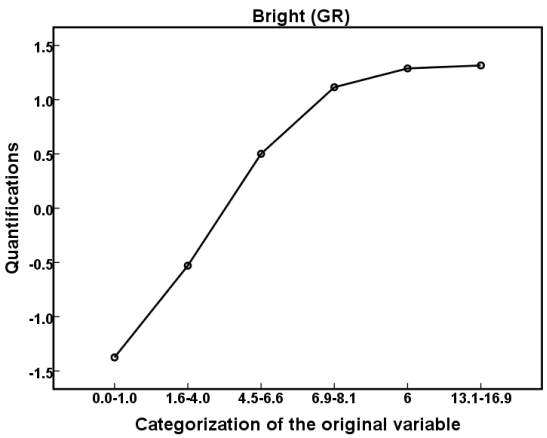

Thin (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
2
2.8-3.1
3.9-7.2
.
12.9-15.1
15.7-18.7
Categorization of the original variable

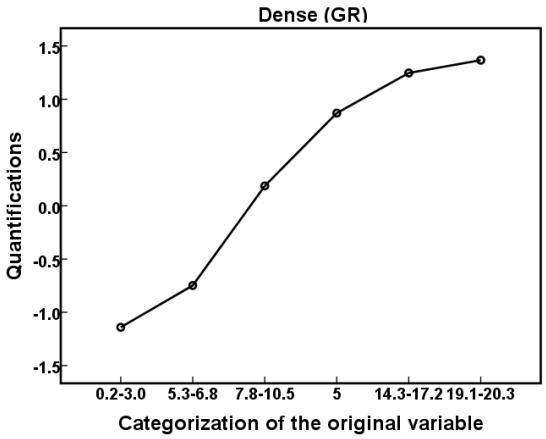

Shrill (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.8-3.6
7.6-13.5
14.8-21.2
5
31.3-40.9
49.9-56.1
Categorization of the original variable

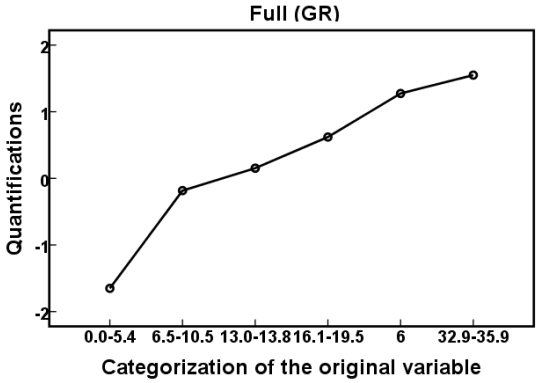

Cold (GR)
Quantifications
2
1
0
-1
-2
0.6-3.0
3.3-5.0
5.9-6.6
7.9-8.2
10.2-11.9
13.8-14.0
Categorization of the original variable

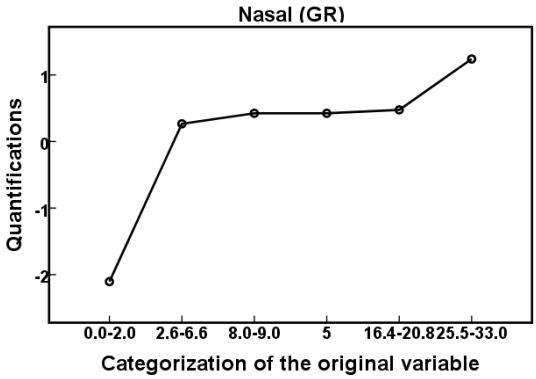

Sharp (GR)
Quantifications
2
1
0
-1
-2
0.0-0.7
2.1-6.6
9.0-10.7
14.6-18.2
6
30.5-36.3
Categorization of the original variable

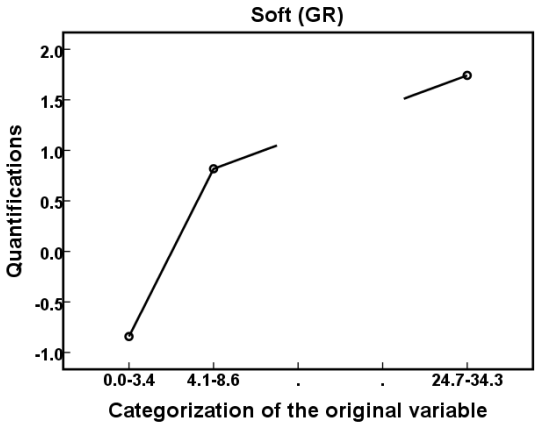

Rich (GR)
Quantifications
3
2
1
0
-1
-2
0.0-1.7
2.8-5.2
5.7-7.6
8.8-10.7
6
19.7-20.5
Categorization of the original variable

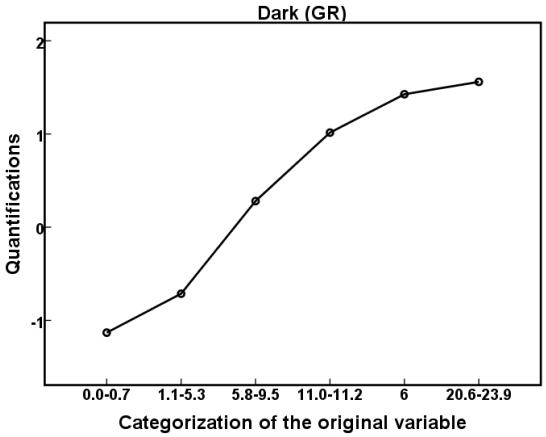

Bright (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0-1.0
1.6-4.0
4.5-6.6
6.9-8.1
6
13.1-16.9
Categorization of the original variable

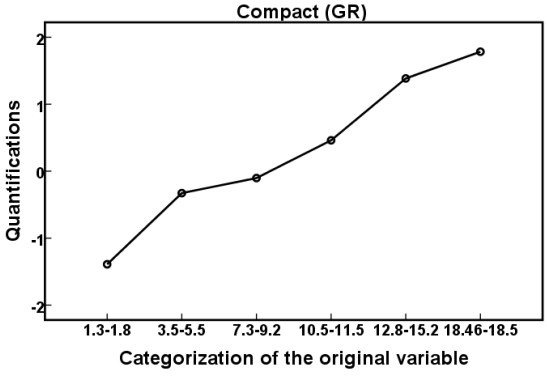

Dense (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.2-3.0
5.3-6.8
7.8-10.5
5
14.3-17.2
19.1-20.3
Categorization of the original variable

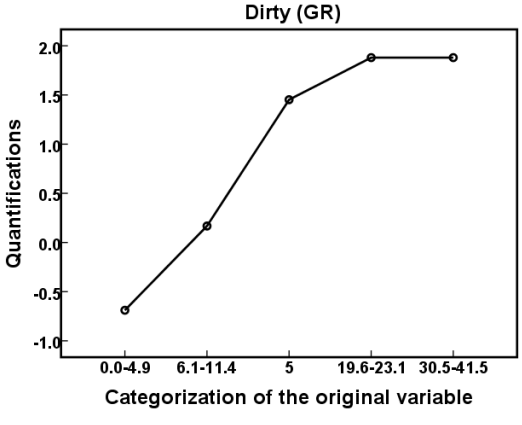

Full (GR)
Quantifications
2
1
0
-1
-2
0.0-5.4
6.5-10.5
13.0-13.8
16.1-19.5
6
32.9-35.9
Categorization of the original variable

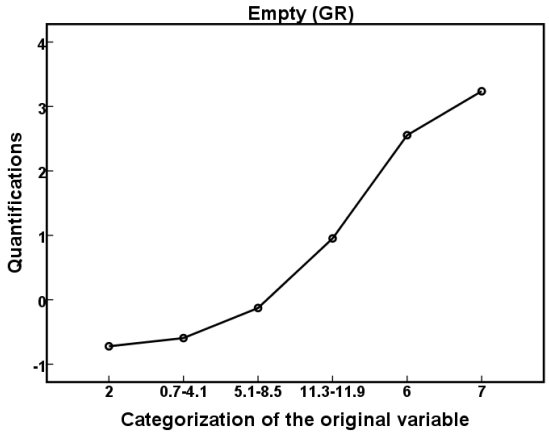

Nasal (GR)
Quantifications
1
0
-1
-2
0.0-2.0
2.6-6.6
8.0-9.0
5
16.4-20.8
25.5-33.0
Categorization of the original variable

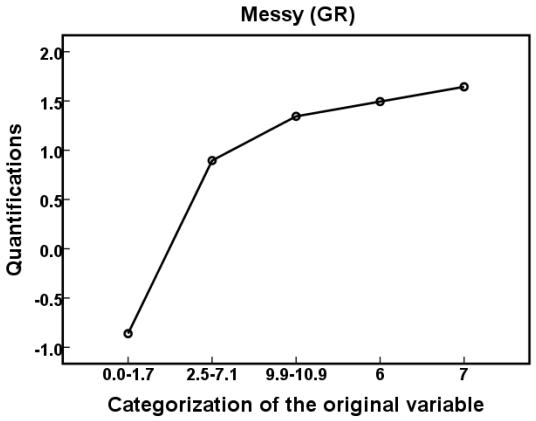

Soft (GR)
Quantifications
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0.0-3.4
4.1-8.6
.
.
24.7-34.3
Categorization of the original variable

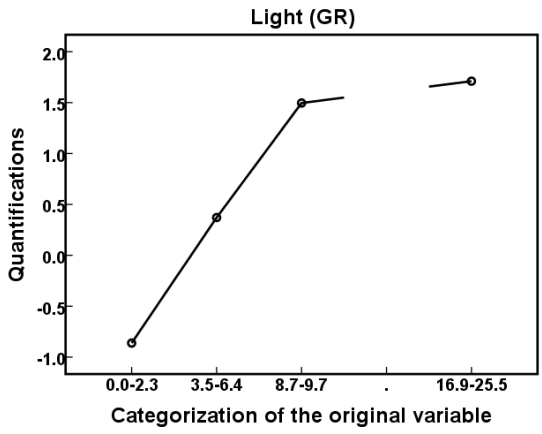

Dark (GR)
Quantifications
2
1
0
-1
0.0-0.7
1.1-5.3
5.8-9.5
11.0-11.2
6
20.6-23.9
Categorization of the original variable

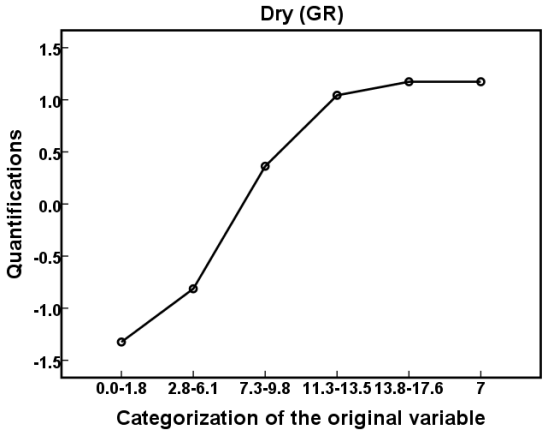

Compact (GR)
Quantifications
2
1
0
-1
-2
1.3-1.8
3.5-5.5
7.3-9.2
10.5-11.5
12.8-15.2
18.46-18.5
Categorization of the original variable

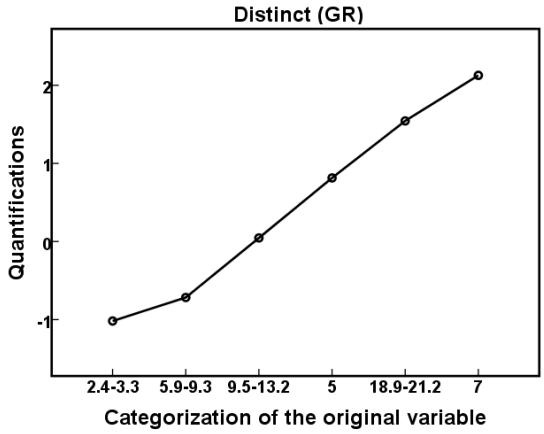

Dirty (GR)
Quantifications
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0.0-4.9
6.1-11.4
5
19.6-23.1
30.5-41.5
Categorization of the original variable

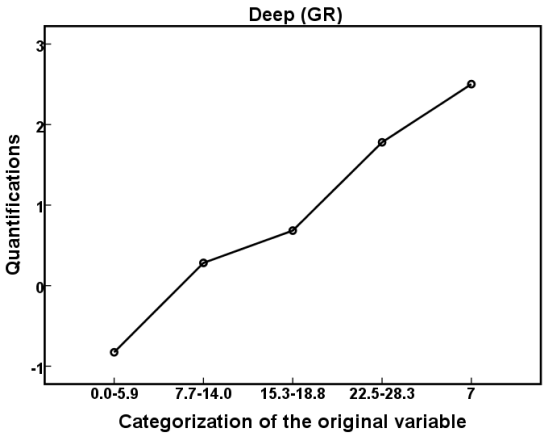

Empty (GR)
Quantifications
4
3
2
1
0
-1
2
0.7-4.1
5.1-8.5
11.3-11.9
6
7
Categorization of the original variable

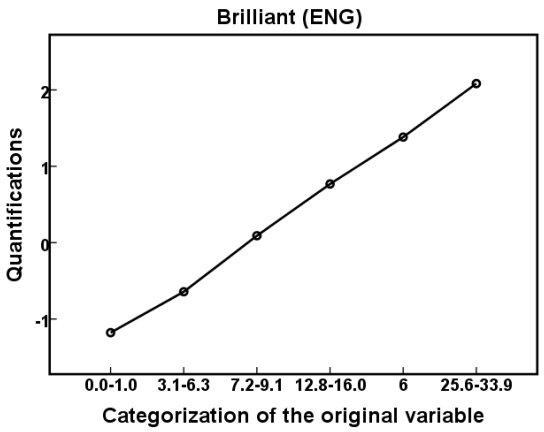

Messy (GR)
Quantifications
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0.0-1.7
2.5-7.1
9.9-10.9
6
7
Categorization of the original variable

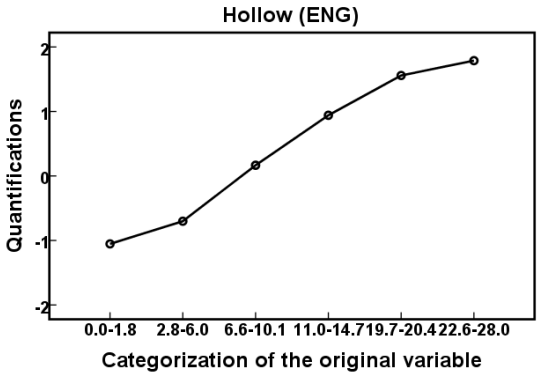

Light (GR)
Quantifications
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0.0-2.3
3.5-6.4
8.7-9.7
.
16.9-25.5
Categorization of the original variable

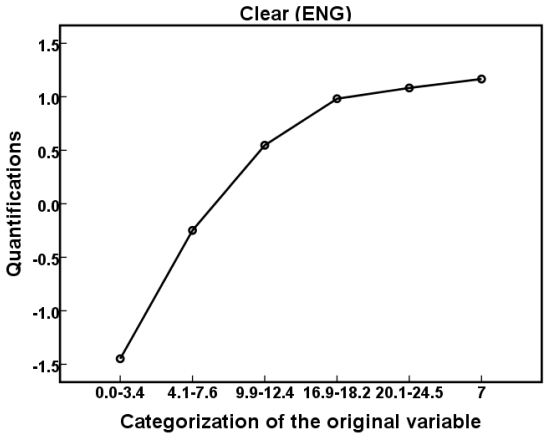

Dry (GR)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0-1.8
2.8-6.1
7.3-9.8
11.3-13.5
13.8-17.6
7
Categorization of the original variable

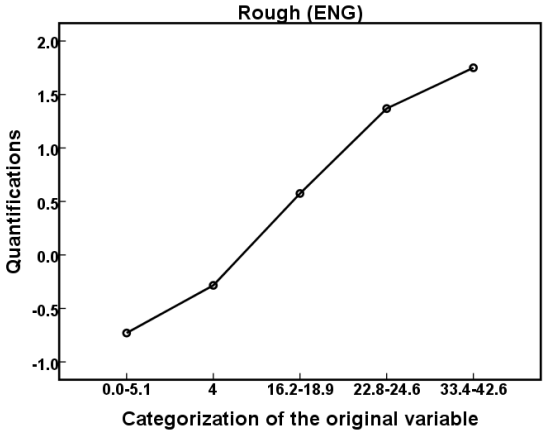

Distinct (GR)
Quantifications
2
1
0
-1
2.4-3.3
5.9-9.3
9.5-13.2
5
18.9-21.2
7
Categorization of the original variable

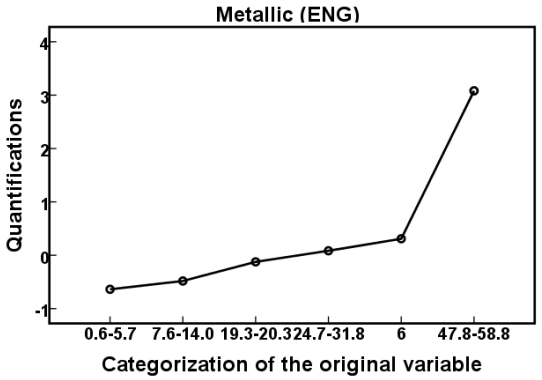

Deep (GR)
Quantifications
3
2
1
0
-1
0.0-5.9
7.7-14.0
15.3-18.8
22.5-28.3
7
Categorization of the original variable

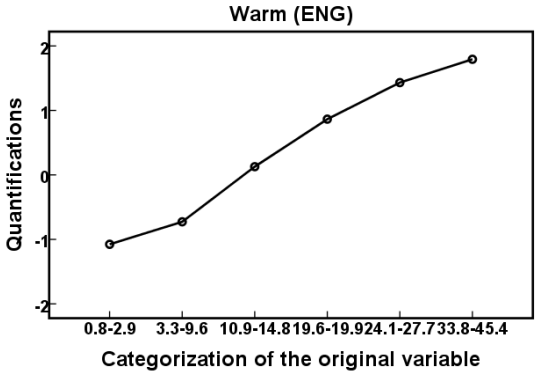

Brilliant (ENG)
Quantifications
2
1
0
-1
0.0-1.0
3.1-6.3
7.2-9.1
12.8-16.0
6
25.6-33.9
Categorization of the original variable

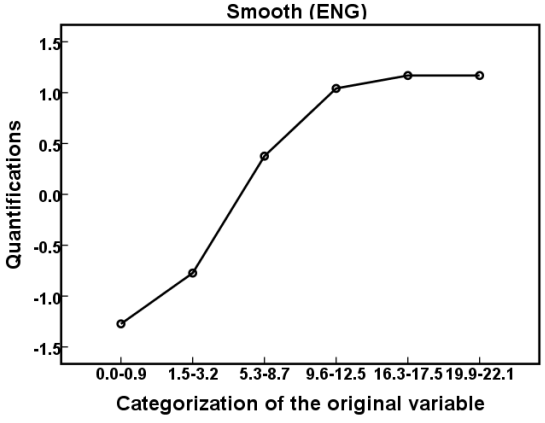

Hollow (ENG)
Quantifications
2
1
0
-1
-2
0.0-1.8
2.8-6.0
6.6-10.1
11.0-14.7
19.7-20.4
22.6-28.0
Categorization of the original variable

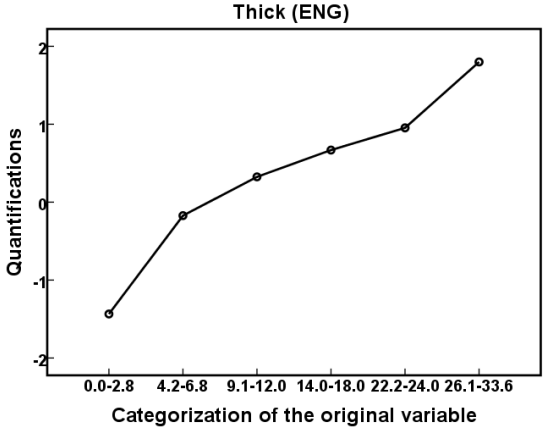

Clear (ENG)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0-3.4
4.1-7.6
9.9-12.4
16.9-18.2
20.1-24.5
7
Categorization of the original variable

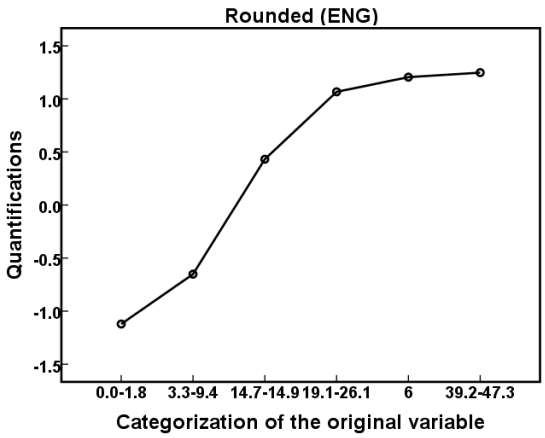

Rough (ENG)
Quantifications
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0.0-5.1
4
16.2-18.9
22.8-24.6
33.4-42.6
Categorization of the original variable

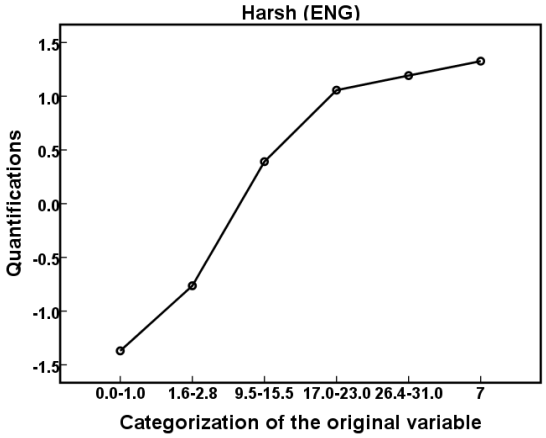

Metallic (ENG)
Quantifications
4
3
2
1
0
-1
0.6-5.7
7.6-14.0
19.3-20.3
24.7-31.8
6
47.8-58.8
Categorization of the original variable

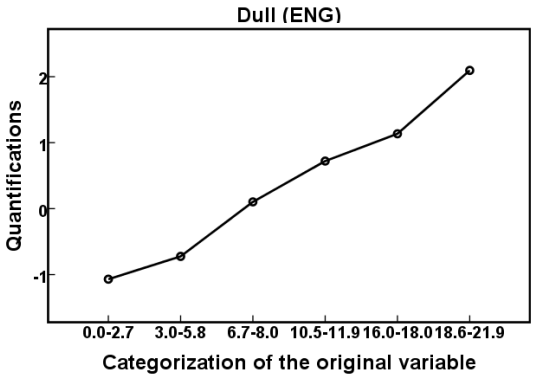

Warm (ENG)
Quantifications
2
1
0
-1
-2
0.8-2.9
3.3-9.6
10.9-14.8
19.6-19.9
24.1-27.7
33.8-45.4
Categorization of the original variable

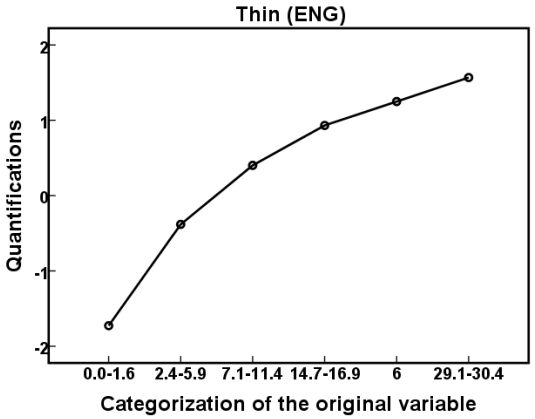

Smooth (ENG)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0-0.9
1.5-3.2
5.3-8.7
9.6-12.5
16.3-17.5
19.9-22.1
Categorization of the original variable

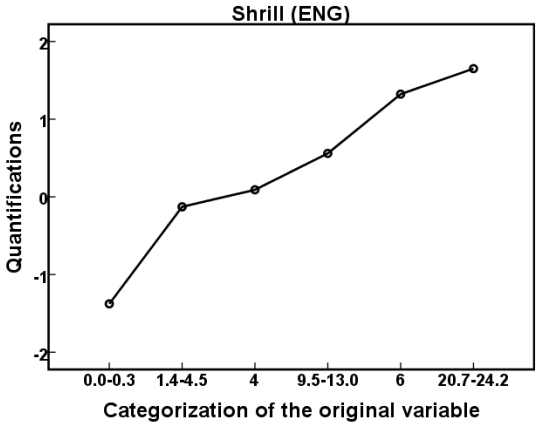

Thick (ENG)
Quantifications
2
1
0
-1
-2
0.0-2.8
4.2-6.8
9.1-12.0
14.0-18.0
22.2-24.0
26.1-33.6
Categorization of the original variable

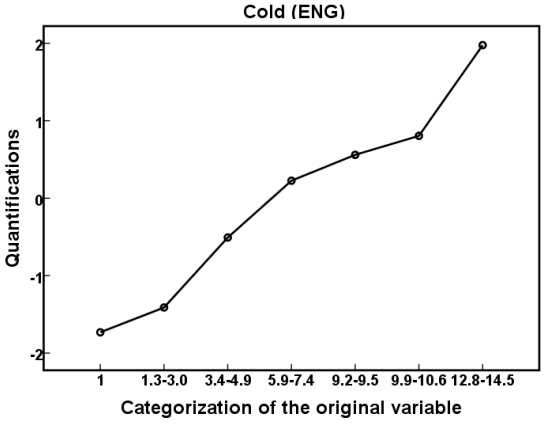

Rounded (ENG)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0-1.8
3.3-9.4
14.7-14.9
19.1-26.1
6
39.2-47.3
Categorization of the original variable

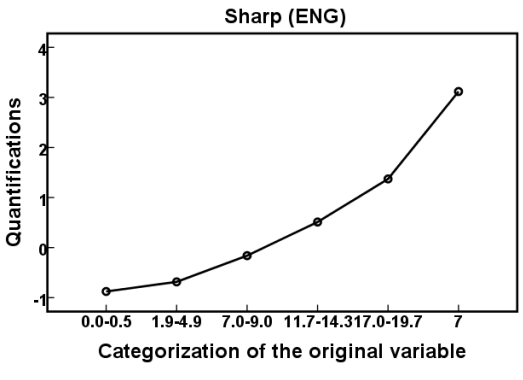

Harsh (ENG)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0-1.0
1.6-2.8
9.5-15.5
17.0-23.0
26.4-31.0
7
Categorization of the original variable

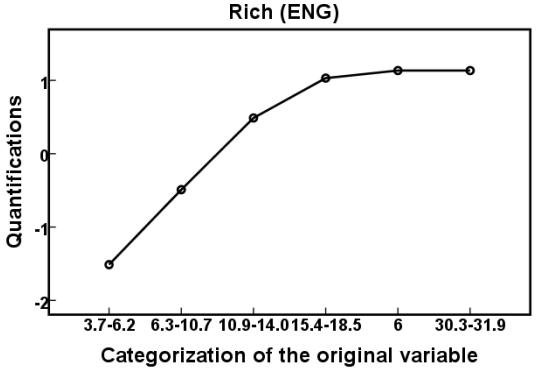

Dull (ENG)
Quantifications
2
1
0
-1
0.0-2.7
3.0-5.8
6.7-8.0
10.5-11.9
16.0-18.0
18.6-21.9
Categorization of the original variable

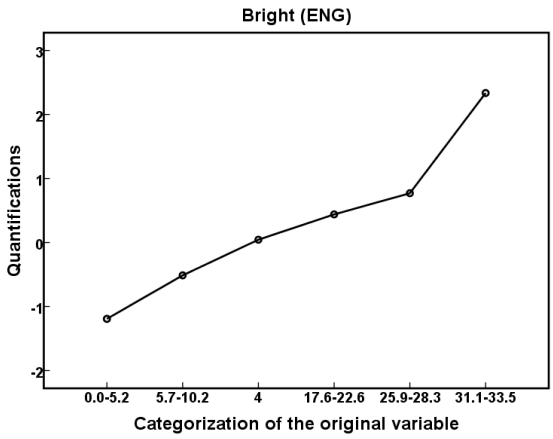

Thin (ENG)
Quantifications
2
1
0
-1
-2
0.0-1.6
2.4-5.9
7.1-11.4
14.7-16.9
6
29.1-30.4
Categorization of the original variable

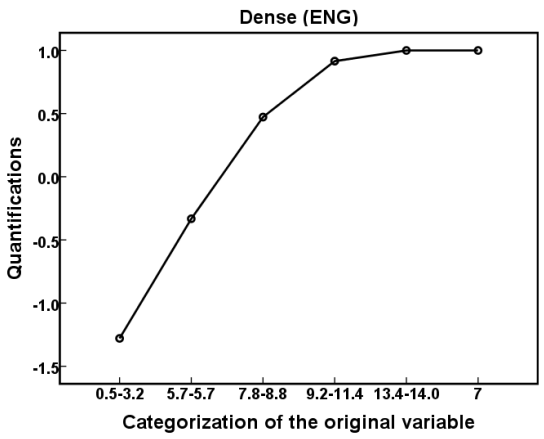

Shrill (ENG)
Quantifications
2
1
0
-1
-2
0.0-0.3
1.4-4.5
4
9.5-13.0
6
20.7-24.2
Categorization of the original variable

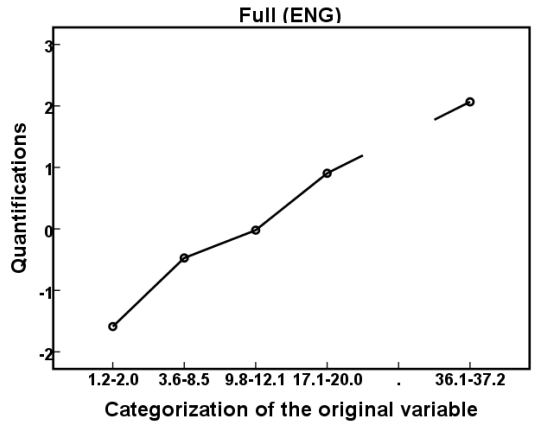

Cold (ENG)
Quantifications
2
1
0
-1
-2
1
1.3-3.0
3.4-4.9
5.9-7.4
9.2-9.5
9.9-10.6
12.8-14.5
Categorization of the original variable

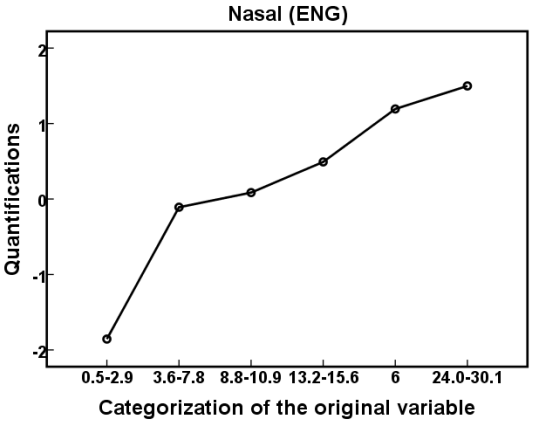

Sharp (ENG)
Quantifications
4
3
2
1
0
-1
0.0-0.5
1.9-4.9
7.0-9.0
11.7-14.3
17.0-19.7
7
Categorization of the original variable

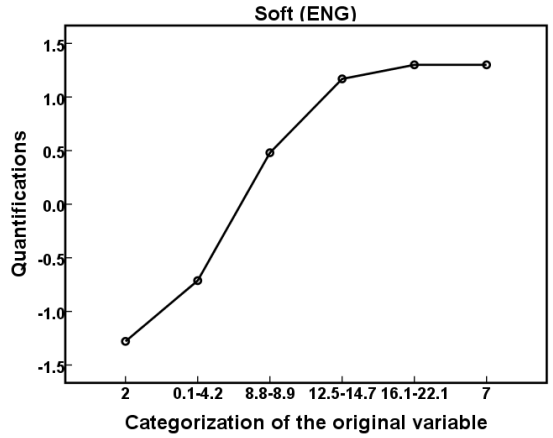

Rich (ENG)
Quantifications
1
0
-1
-2
3.7-6.2
6.3-10.7
10.9-14.0
15.4-18.5
6
30.3-31.9
Categorization of the original variable

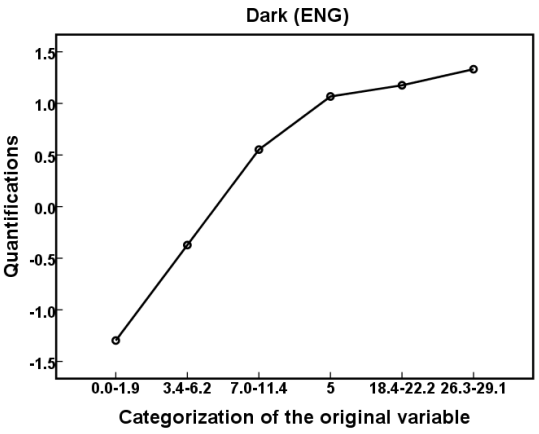

Bright (ENG)
Quantifications
3
2
1
0
-1
-2
0.0-5.2
5.7-10.2
4
17.6-22.6
25.9-28.3
31.1-33.5
Categorization of the original variable

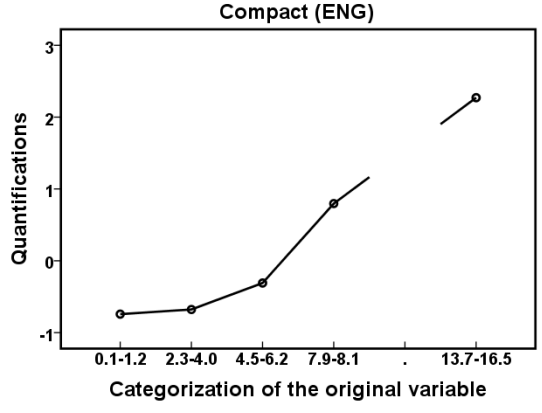

Dense (ENG)
Quantifications
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.5-3.2
5.7-5.7
7.8-8.8
9.2-11.4
13.4-14.0
7
Categorization of the original variable

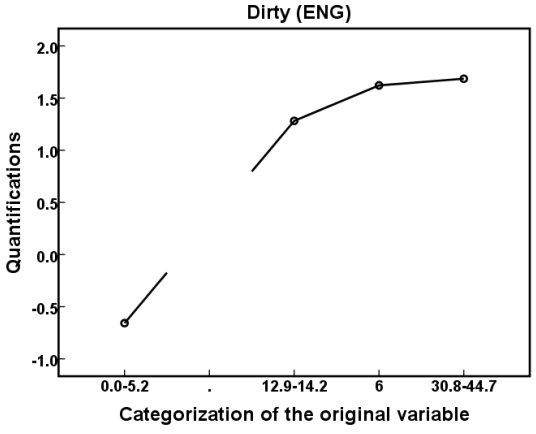

Full (ENG)
Quantifications
3
2
1
0
-1
-2
1.2-2.0
3.6-8.5
9.8-12.1
17.1-20.0
.
36.1-37.2
Categorization of the original variable


Nasal (ENG)
Quantifications
2
1
0
-1
-2
0.5-2.9
3.6-7.8
8.8-10.9
13.2-15.6
6
24.0-30.1
Categorization of the original variable


Soft (ENG)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
2
0.1-4.2
8.8-8.9
12.5-14.7
16.1-22.1
7
Categorization of the original variable


Dark (ENG)
Quantifications
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0-1.9
3.4-6.2
7.0-11.4
5
18.4-22.2
26.3-29.1
Categorization of the original variable


Compact (ENG)
Quantifications
3
2
1
0
-1
0.1-1.2
2.3-4.0
4.5-6.2
7.9-8.1
.
13.7-16.5
Categorization of the original variable


Dirty (ENG)
Quantifications
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0.0-5.2
.
12.9-14.2
6
30.8-44.7
Categorization of the original variable

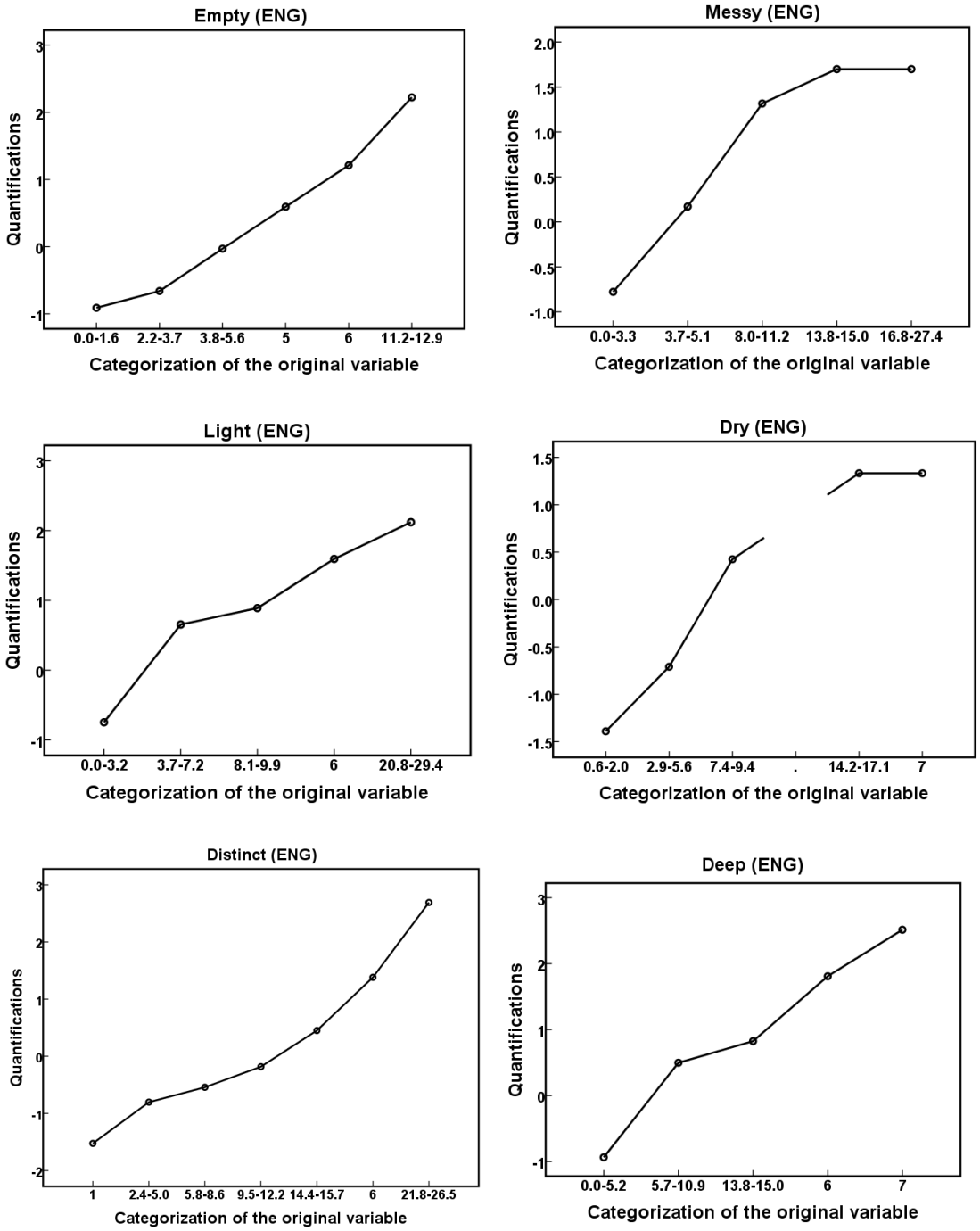


Figure A.1: Transformation plots corresponding to the 30 adjectives for both Greek and English.

Bibliography

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

V. Alluri, P. Toiviainen, I. P. Jääskeläinen, E. Glerean, M. Sams, and E. Brattico. Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, 59(4):3677–3689, 2012.

X. Amatriain, J. Bonada, A. Loscos, and X. Serra. Spectral processing. In Udo Zölzer, editor, *DAFX - Digital Audio Effects*, pages 373–438. John Wiley and Sons Ltd, Chichester, England, 2002.

ANSI. American National Standard–Psychoacoustical Terminology. Technical report, (American National Standards Institute, New York), 1973.

M. Barthet, P. Depalle, R. Kronland-Martinet, and S. Ystad. Acoustical correlates of timbre and expressiveness in clarinet performance. *Music Perception*, 28(2):135–153, 2010a.

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

M. Barthet, P. Depalle, R. Kronland-Martinet, and S. Ystad. Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance. *Music Perception*, 28(3):265–278, 2011.

L. Boroditsky, L. A. Schmidt, and W. Phillips. Sex, syntax, and semantics. In D. Genter and S. Goldin-Meadow, editors, *Language in Mind*. Cambridge, MA: MIT Press, 2003.

A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1994.

A. Burgess. *A Clockwork Orange*. Norton paperback fiction. W. W. Norton, 1986. ISBN 9780393312836.

J. A. Burgoyne and S. McAdams. Non-linear scaling techniques for uncovering the perceptual dimension of timbre. In *Proceedings of the International Computer Music Conference*, volume 1, pages 73 – 76, 2007.

J. A. Burgoyne and S. McAdams. A meta-analysis of timbre perception using nonlinear extensions to CLASCAL. In R. Kronland-Martinet, S. Ystad, and J. Kristoffer, editors, *Computer Music Modeling and Retrieval. Sense of Sounds*, pages 181–202. Springer, 2008.

A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg. Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, 118(1):471–482, 2005.

A. Caclin, E. Brattico, M. Tervaniemi, R. Näätänen, D. Morlet, M.-H. Giard, and S. McAdams. Separate neural processing of timbre dimensions in auditory sensory memory. *Journal of Cognitive Neuroscience*, 18(12):1959–1972, 2006.

A. Caclin, M-H. Giard, B. K. Smith, and S. McAdams. Interactive processing of timbre dimensions: A Garner interference study. *Brain Research*, 1138:159–170, 2007.

M. Caetano, J. J. Burred, and X. Rodet. Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, pages 11–21, 2010.

M. Campbell, C. Greated, and A. Myers. *Musical Instruments: History, Technology and Performance of Instruments of Western Music*. Oxford University Press, USA, 2006.

J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

J. J. F. Commandeur and W. J. Heiser. Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices. Technical report, Leiden: Department of Data Theory, University of Leiden, 1993.

A. L. Comrey and H. B. Lee. *A First Course in Factor Analysis*. Psychology Press, 2nd edition, 1992.

A. P. M. Coxon and C. Jones. Multidimensional scaling: Exploration to confirmation. *Quality and Quantity*, 14(1):31–73, 1980.

J. De Leeuw, F. W. Young, and Y. Takane. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41(4):471–503, 1976.

A. C. Disley and D. M. Howard. Spectral correlates of timbral semantics relating to the pipe organ. In *Speech, Music and Hearing (TMH-QPSR)*, volume 46, pages 25–40, 2004.

A. C. Disley, D. M. Howard, and A. D. Hunt. Timbral description of musical instruments. In *Proceedings of the 9th International Conference on Music Perception and Cognition (ICMPC 09)*, pages 61–68, 2006.

S. Donnadieu. Mental representation of the timbre of complex sounds. In J. W. Beauchamp, editor, *Analysis, Synthesis, and Perception of Musical Sounds*, pages 272–319. Springer, New York, USA, 2007.

J. C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1): 95–104, 1974.

D. E. Ehresman and D. L Wessel. Perception of timbral analogies. Technical report, IRCAM 13/78, (IRCAM, Centre Georges Pompidou, Paris), 1978.

T. M. Elliott, L. S. Hamilton, and F. E. Theunissen. Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *Journal of the Acoustical Society of America*, 133(1):389–404, 2012.

R. Ethington and B. Punch. Seawave: A system for musical timbre description. *Computer Music Journal*, 18(1):30–39, 1994.

L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272–299, 1999.

H. Fastl and E. Zwicker. *Psychoacoustics, Facts and Models*. Springer, 2007.

A. Faure, S. McAdams, and V. Nosulenko. Verbal correlates of perceptual dimensions of timbre. In *Proceedings of the 4th International Conference on Music Perception and Cognition (ICMPC 04)*, pages 79–84, 1996.

C. Fritz, A. F. Blackwell, I. Cross, J. Woodhouse, and B. C. J. Moore. Exploring violin sound quality: Investigating English timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties. *Journal of the Acoustical Society of America*, 131(1): 783–794, 2012.

D. George and P. Mallery. *SPSS for Windows step by step: A simple guide and reference 11.0 Update*. Allyn & Bacon, Boston, 4th edition, 2003.

B. L. Giordano and S. McAdams. Sound source mechanics and musical timbre perception: Evidence from previous studies. *Music Perception*, 28(2):157–170, 2010.

B. L. Giordano, D. Rocchesso, and S. McAdams. Integration of acoustical information in the perception of impacted sound sources: The role of information accuracy and exploitability. *Journal of Experimental Psychology: Human Perception and Performance*, 36(2):462–479, 2010.

B. L. Giordano, S. McAdams, R. J. Zatorre, N. Kriegeskorte, and P. Belin. Abstract encoding of auditory objects in cortical activity patterns. *Cerebral Cortex*, 2012.

P. E. Green, F. Carmone, and S. M. Smith. *Multidimensional Scaling: Concept and Applications*. Allyn and Bacon, Boston, 1989.

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

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

J. M. Hajda. A new model for segmenting the envelope of musical signals: The relative salience of steady state versus attack, revisited. In *101st Audio Engineering Society Convention*, Los Angeles, 1996.

J. M. Hajda. Relevant acoustical cues in the identification of western orchestral instrument tones. *(abstract), The Journal of the Acoustical Society of America*, 102(5):3085, 1997.

J. M. Hajda. *The Effect of Time-Variant Acoustical Properties on Orchestral Instrument Timbres*. PhD thesis, University of California, Los Angeles, 1999.

J. M. Hajda. The effect of dynamic acoustical features on musical timbre. In J. W. Beauchamp, editor, *Analysis, Synthesis, and Perception of Musical Sounds*, pages 250–271. Springer, New York, USA, 2007.

J. M. Hajda, R. A. Kendall, E. C. Carterette, and M. L. Harshberger. Methodological issues in timbre research. In I. Deliège and J. Sloboda, editors, *Perception and Cognition of Music*, pages 253–306. (Psychology Press, Hove, UK), 1997.

S. Handel and M. L. Erickson. Sound source identification: The possible role of timbre transformations. *Music Perception*, 21(4):587–610, 2004.

H. H. Harman. *Modern Factor Analysis*. University of Chicago Press, Chicago, IL, 3rd edition, 1976.

C. Hourdin, G. Charbonneau, and T. Moussa. A multidimensional scaling analysis of musical instrument's time varying spectra. *Computer Music Journal*, 21(2):40–55, 1997.

D. Howard and A. Tyrrell. Psychoacoustically informed spectrography and timbre. *Organised Sound*, 2(2):65–76, 1997.

D. Howard, A. Disley, and A. Hunt. Timbral adjectives for the control of a music synthesizer. In *19th International Congress on Acoustics*, 2007.

G. Hutcheson and G. Sofroniou. *The Multivariate Social Scientist*. SAGE Publications Ltd, London, 1999.

IBM SPSS Statistics 20 Algorithms. IBM Corp., 2011. `ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf`.

P. Iverson and C. L. Krumhansl. Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, 94(5):2595–2603, 1993.

R. I. Jennrich and P. F. Sampson. Rotation for simple loadings. *Psychometrika*, 31(3), 1966.

K. Jensen. *Timbre Models of Musical Sounds*. PhD thesis, University of Copenhagen, Denmark, 1999.

V. Karbusicky. *Grundriß der musikalischen semantik*. Wissenschaftliche Buchgesellschaft, Darmstadt, 1986.

Kawachi. How impression and recognition are processed in the cerebrum. In *KANSEI / Human / Computer*, pages 78–120. Fujitsu Books (in Japanese), 1995.

R. A. Kendall and E. C. Carterette. Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, 8(4):369–404, 1991.

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

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

R. A. Kendall and E. C. Carterette. Difference thresholds for timbre related to spectral centroid. In *Proceedings of the 4th International Conference on Music Perception and Cognition (ICMPC 04)*, pages 91–95, 1996.

R. A. Kendall, E. C. Carterette, and J. M. Hajda. Perceptual and acoustical features of natural and synthetic orchestral instrument tones. *Music Perception*, 16(3):327–364, 1999.

P. Kline. *The handbook of psychological testing*. Routledge, London, 2nd edition, 1999.

L. Knopoff. An index for the relative quality among musical instruments. *Ethnomusicology*, 7 (3):229–233, 1963.

S. Koelsch. Towards a neural basis of processing musical semantics. *Physics of Life Reviews*, 8 (2):89–105, 2011.

J. Krimphoff. Analyse acoustique et perception du timbre. Master's thesis, Université du Maine, Le Mans, France, 1993.

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

C. L. Krumhansl. Why is musical timbre so hard to understand? In *Structure and Perception of Electroacoustic Sound and Music: Proceedings of the Marcus Wallenberg Symposium*, pages 43–53, 1989.

C. L. Krumhansl and P. Iverson. Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance*, 18(3):739–751, 1992.

J. B. Kruskal. Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–28, 1964a.

J. B. Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2): 115–130, 1964b.

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

O. Lartillot, P. Toiviainen, and T. Eerola. A matlab toolbox for music information retrieval. In *Data Analysis, Machine Learning and Applications*, pages 261–268. Springer, Berlin Heidelberg, 2008.

A. Lehrer. Can wines be brawny?: Reflections on wine vocabulary. In Barry C. S., editor, *Questions of Taste: The Philosophy of Wine*, pages 127–140. Signal Books, Oxford, 2007.

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

T. Lokki, J. Pätynen, A. Kuusinen, H. Vertanen, and S. Tervo. Concert hall acoustics assessment with individually elicited attributes. *Journal of the Acoustical Society of America*, 130(2): 835–849, 2011.

D. A. Luce. *Physical correlates of nonpercussive musical instrument tones*. PhD thesis, Massachusetts Institute of Technology, 1963.

I. MacDonald. *Revolution in the Head: The Beatles' Records and the Sixties*. Pimlico, 1995.

J. Marozeau and A. de Cheveignè. The effect of fundamental frequency on the brightness dimension of timbre. *Journal of the Acoustical Society of America*, 121(1):383–387, 2007.

J. Marozeau, A. de Cheveignè, S. McAdams, and S. Winsberg. The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America*, 114(5):2946–2957, 2003.

S. McAdams. Perspectives on the contribution of timbre to musical structure. *Computer music journal*, 23(3):85–102, 1999.

S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192, 1995.

S. McAdams, J. W. Beauchamp, and S. Meneguzzi. Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters. *Journal of the Acoustical Society of America*, 105(2):882–897, 1999.

J. J. Meulman and W. J. Heiser. *PASW Categories 18, Chapter 3*. SPSS Inc., Chicago, 2008.

J. R. Miller and E. C. Carterette. Perceptual space for musical structures. *Journal of the Acoustical Society of America*, 58(3):711–720, 1975.

B. C. J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, San Diego, 4th edition, 2003.

B. C. J. Moore, B. R. Glasberg, and B. Thomas. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240, 1997.

J. A. Moorer. Signal processing aspects of computer music–a survey. *Computer Music Journal*, 1(1):4–37, 1977.

O. Moravec and J. Štěpánek. Verbal description of musical sound timbre in Czech language. In *Proceedings of the Stockholm Music Acoustics Conference (SMAC03)*, pages 643–645, 2003.

I. Morlini and S. Zani. Dissimilarity and similarity measures for comparing dendrograms and their applications. *Advances in Data Analysis and Classification*, 6(2):85–105, 2012.

T. Neher, T. Brookes, and F. Rumsey. A hybrid technique for validating unidimensionality of perceived variation in a spatial auditory stimulus set. *Journal of the Audio Engineering Society*, 54(4):259–275, 2006.

G. A. Nicol. *Development and Exploration of a Timbre Space Representation of Audio*. PhD thesis, University of Glasgow, 2005.

A. Nykänen, Ö. Johansson, J. Lundberg, and J. Berg. Modelling perceptual dimensions of saxophone sounds. *Acta Acustica united with Acustica*, 95(3):539–549, 2009.

G. S. Ohm. Über die definition des tones, nebst daran geknüpfter theorie der sirene und ähnlicher tonbildender vorrichtungen. *Annalen der Physik*, 135(8):513–565, 1843. (See discussion in Wever, 1949).

F. Opolko and J. Wapnick. *McGill University Master Samples collection on DVD*. McGill University, Montréal, Québec, Canada, 2006.

C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, Urbana, IL, 1957.

J. G. Painter and S. Koelsch. Can out-of-context musical sounds convey meaning? An ERP study on the processing of meaning in music. *Psychophysiology*, 48(5):645–655, 2011.

G. Papanikolaou and C. Pastiadis. Multiple dichotomies in timbre research. *Archives of Acoustics*, 34(2):137–155, 2009.

G. Peeters. A large set of audio features for sound description (similarity and classification) in CUIDADO project. Technical report, CUIDADO 1st Project Report, IRCAM, Analysis/Synthesis Team, 2004. pp. 1–25.

G. Peeters, S. McAdams, and P. Herrera. Instrument sound description in the context of MPEG-7. In *Proceedings of the 2000 International Computer Music Conference*, pages 166–169, 2000.

G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams. The Timbre Toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, 130(5):2902–2916, 2011.

C. S. Peirce. *The collected papers of Charles Sanders Pierce*. Harvard University Press, Cambridge MA, 1931/1958.

R. Plomp. Timbre as a multidimensional attribute of complex tones. In R. R. Plomp and G. F. Smoorenburg, editors, *Frequency Analysis and Periodicity Detection in Hearing*, pages 397–414. Sijthoff, Leiden, 1970.

R. Plomp. Timbre of complex tones. In R. Plomp, editor, *Aspects of Tone Sensation: A Psychophysical Study*, pages 85–110. Academic Press, London, 1976.

H. F. Pollard and E. V. Jansson. A tristimulus method for the specification of musical timbre. *Acustica*, 51(3):162–171, 1982.

R. L. Pratt and P. E. Doak. A subjective rating scale for timbre. *Journal of Sound and Vibration*, 45(3):317–328, 1976.

J. O. Ramsay. Is multidimensional scaling magic or science? *Contemporary Psychology*, 33(10): 874–875, 1988.

C. Roads, editor. *The Computer Music Tutorial*. MIT Press, Cambridge, Massachusetts, 1996.

C. Romesburg. *Cluster Analysis for Researchers*. Lulu Press, North Carolina, 2004.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20(1):53–65, 1987.

C. Saitis, B. L. Giordano, C. Fritz, and G. P. Scavone. Perceptual evaluation of violins: A quantitative analysis of preference judgments by experienced players. *Journal of the Acoustical Society of America*, 132(6):4002–4012, 2012.

E. Samoylenko, S. McAdams, and V. Nosulenko. Systematic analysis of verbalizations produced in comparing musical timbres. *International Journal of Psychology*, 31(6):255–278, 1996.

G. J. Sandell. Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration. *Music Perception*, 13(2):209–246, 1995.

M. Sankiewicz and G. Budzynski. Reflections on sound timbre definitions. *Archives of Acoustics*, 32(3):442–452, 2007.

P. Schaeffer. *Traité des objets musicaux [Treatise on musical objects]*. Seuil, Paris, 1966.

E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1331–1334, 1997.

A. Schoenberg. *Harmonielehre*. University of California Press, Berkeley Los Angeles, 3rd edition, 1922.

E. Schubert and J. Wolfe. Does timbral brightness scale with frequency and spectral centroid? *Acta Acustica united with Acustica*, 92(5):820–825, 2006.

E. Schubert, J. Wolfe, and A. Tarnopolsky. Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of 8th International Conference on Music Perception and Cognition (ICMPC 08)*, pages 654–657, 2004.

X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S. T. Pope, A. Piccialli, and G. de Poli, editors, *Musical Signal Processing*, pages 91–122. Swets & Zeitlinger publishers, 1997.

X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1990.

R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function: I. *Psychometrika*, 27(2):125–140, 1962a.

R. N. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function: II. *Psychometrika*, 27(3):219–246, 1962b.

R. N. Shepard. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3(2): 287–315, 1966.

D. Smalley. Spectromorphology: explaining sound-shapes. *Organised Sound*, 2(2):107–126, 1997.

F. G. Stremler. Information and digital transmissions. In *Introduction to communication systems*, pages 506–507. Addison-Wesley, 3rd edition, 1990.

B. G. Tabachnick and L. S. Fidell. *Using Multivariate Statistics*. Allyn & Bacon, 5th edition, 2006.

Yoshio Takane, Forrest W. Young, and Jan De Leeuw. Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42(1):7–67, 1977.

J. B. Tennenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

H. Terasawa. *A Hybrid Model for Timbre Perception: Quantitative Representations of Sound Colour and Density*. PhD thesis, Stanford University, 2009.

C. Traube, M. Bernays, and M. Bellemare. Perception, verbal description and gestural control of piano timbre. *The Journal of the Acoustical Society of America*, 123(5):3657–3657, 2008.

J. Štěpánek. Musical sound timbre: Verbal descriptions and dimensions. In *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pages 121–126, 2006.

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

G. von Bismarck. Sharpness as an attribute of the timbre of steady sounds. *Acustica*, 30:159–172, 1974b.

H. L. F. von Helmholtz. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. New York: Dover (1954), 4th edition, 1877. English translation by A. J. Ellis.

E. M. von Hornbostel and E. Sachs. Systematik der Musikinstrumente: Ein Versuch. translated as "Classification of Musical Instruments" by Anthony Baines and Klaus Wachsmann. *Galpin Society Journal (1961)*, 14(4/5):3–29, 1914.

A. Vurma, M. Raju, and A. Kuuda. Does timbre affect pitch?: Estimations by musicians and non-musicians. *Psychology of Music*, 39(3):291–306, 2010.

S. Wake and T. Asahi. Sound retrieval with intuitive verbal expressions. In *International Conference on Auditory Display (ICAD'98)*, pages 1–5, 1998.

S. L. Weinberg and V. C. Menil. The recovery of structure in linear and ordinal data: INDSCAL versus ALDSCAL. *Multivariate Behavioral Research*, 28(2):215–233, 1993.

D. L. Wessel. Psychoacoustics and music: A report from Michigan State University. *PACE: Bulletin of the Computer Arts Society*, 30:1–2, 1973.

D. L. Wessel. Timbre space as a musical control structure. *Computer Music Journal*, 3(2):45–52, 1979.

D. Williams and T. Brookes. Perceptually motivated audio morphing: Brightness. In *Proceedings of the 122nd Audio Engineering Convention*, 2007.

D. Williams and T. Brookes. Perceptually-motivated audio morphing: Warmth. In *Proceedings of the 128th Audio Engineering Society Convention*, 2010.

S. Winsberg and J. D. Carroll. A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model. *Psychometrika*, 54(2):217–229, 1989.

S. Winsberg and G. De Soete. A latent class approach to fitting the weighted Euclidean model: CLASCAL. *Psychometrika*, 58(2):315–330, 1993.

J. A. Woodward and J. E. Overall. Factor analysis of rank-ordered data: An old approach revisited. *Psychological Bulletin*, 83(5):864–867, 1976.

F. W. Young. Nonmetric multidimensional scaling: Recovery of metric information. *Psychometrica*, 35(4):455–473, 1970.

F. W. Young, Y. Takane, and R. Lewyckyj. ALSCAL: A nonmetric multidimensional scaling program with several different options. *Behavioral Research Methods and Instrumentation*, 10(3):451–453, 1978.

A. Zacharakis and J. D. Reiss. An additive synthesis technique for independent modification of the auditory perceptions of brightness and warmth. In *Proceedings of the 130th Audio Engineering Society Convention*, 2011.

A. Zacharakis, K. Pastiadis, G. Papadelis, and J. D. Reiss. An investigation of musical timbre: Uncovering salient semantic descriptors and perceptual dimensions. In *Proceedings of the International Society of Music Information Retrieval Conference (ISMIR 12)*, pages 807–812, 2011.

A. Zacharakis, K. Pastiadis, J. D. Reiss, and G. Papadelis. Analysis of musical timbre semantics through metric and non-metric data reduction techniques. In *Proceedings of the 12th International Conference on Music Perception and Cognition (ICMPC 12) and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM 08)*, pages 1177–1182, 2012.

A. Zacharakis, K. Pastiadis, and J. D. Reiss. An inter-language study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception*, accepted.

M. Zaunschirm, J. D. Reiss, and A. Klapuri. A sub-band approach to modification of musical transients. *Computer Music Journal*, 36(2):23–36, 2012.