
An Exploration of Sound Timbre Using Perceptual and Time-Varying Frequency Spectrum Techniques

David Paul Creasey

BEng(Hons), AMIEE

**Thesis Submission in Partial Fulfilment of the Requirements for the
Degree of Doctor of Philosophy in Electronics**

**University of York,
Department of Electronics,
United Kingdom.**

Supported by EPSRC Award 94315368

Supervisors: Prof.A.M.Tyrrell, Prof.D.M.Howard

March 1998

THE UNIVERSITY *of York*



Hypothesis

Timbre space is a perceptually-structured complex multi-dimensional form, which is ineffectively and imprecisely characterised by a low number of exclusive single dimensions or general descriptions, but can be represented by grouped features derived from time and frequency domain representations.

Abstract

This thesis describes the investigation of sound timbre using perceptual and acoustical techniques, with 153 input stimuli. The acoustical methods are based on time and frequency domain representations. The thesis covers the following areas of work:

1. A consideration of previous research in timbre, the different structural forms associated with it, and different definitions concerning timbre and the timbre space representation.
2. A study concerning perceptual similarity reactions to the input stimuli, a statistical analysis of the result structure, and the implications for understanding of the structure of timbral audition.
3. Analysis and synthesis using a time-varying frequency spectrum model, with adaptive viewpoint properties to achieve appropriate time-frequency resolution.
4. Extraction of 335 timbral features from the spectral form, a statistical analysis to find those features which describe perceptual differences between stimuli, and an investigation of timbral dimensionality.

Contents

| | |
|--|-----------|
| Hypothesis | 2 |
| Abstract | 3 |
| Acknowledgements | 13 |
| Declaration | 15 |
| 1 Introduction | 16 |
| 1.1 Background to Research in Timbre | 16 |
| 1.2 Motivation and Aims | 17 |
| 1.3 Details of Hypothesis | 19 |
| 1.4 Contributions to Knowledge Made by This Research | 20 |
| 1.5 Ambiguous Terms | 21 |
| 2 Perspectives and Research into Timbre | 23 |

| | | |
|----------|--|-----------|
| 2.1 | Introduction | 23 |
| 2.2 | Definitions of Timbre : Overview | 24 |
| 2.3 | Definitions of Timbre : Specific Characteristics | 28 |
| 2.4 | Definitions for this Research | 38 |
| 2.5 | Quantities and Qualities of Input Stimuli | 41 |
| 2.6 | Use and Application Concepts | 44 |
| 2.7 | Dimensionality and Model Structure Concepts | 46 |
| 2.8 | Overview of Research Techniques | 58 |
| 2.9 | Important Aspects of the Spectral Form | 66 |
| 2.10 | Specific Features of the Spectral Form | 71 |
| 2.11 | Spectral Correlates of Timbral Semantics | 76 |
| 2.12 | Limitations of Information | 81 |
| 2.13 | Conclusions | 87 |
| 3 | Perceptual Study | 88 |
| 3.1 | Introduction | 88 |
| 3.2 | The Reasons for the Perceptual Study | 89 |
| 3.3 | Mechanisms of Timbre Perception | 91 |
| 3.4 | Perceptual Study Technique | 101 |
| 3.5 | Analysis of Results | 108 |
| 3.6 | Limitations of the Experiment | 141 |
| 3.7 | Conclusions | 143 |

| | | |
|--------------|--|----------------|
| 4 | Analysis-Synthesis Model | 146 |
| 4.1 | Introduction | 146 |
| 4.2 | Background to Spectral Systems | 147 |
| 4.3 | Previous Time-Varying Frequency Spectrum Systems | 149 |
| 4.4 | Analysis-Synthesis Technique Overview | 152 |
| 4.5 | Method and Results | 154 |
| 4.6 | Conclusions | 169 |
| 5 | Timbral Feature Extraction and Analysis | 170 |
| 5.1 | Introduction | 170 |
| 5.2 | Background to Feature Extraction and Analysis | 171 |
| 5.3 | Major Elements of Feature Extraction | 173 |
| 5.4 | Details of Feature Extraction | 177 |
| 5.5 | Overview of Feature Set Analysis | 196 |
| 5.6 | Details of Feature Set Analysis | 202 |
| 5.7 | Discussion and Context of Results | 224 |
| 5.8 | Conclusions | 225 |
| 6 | Conclusions | 228 |
| 6.1 | Introduction | 228 |
| 6.2 | Summary of Chapters | 229 |
| 6.3 | Novel Aspects | 234 |
| 6.4 | Potential Further Work | 236 |

| | | |
|----------|--|------------|
| 6.5 | Confirmation of Hypothesis | 238 |
| 6.6 | Concluding Remarks | 240 |
| A | Descriptions of Sound Samples | 244 |
| B | Mathematics Relevant to This Research | 251 |
| B.1 | Basic Statistical Methods | 252 |
| B.2 | Multivariate Scaling Algorithms | 254 |
| B.3 | Group Discrimination Algorithms | 257 |
| | Glossary | 261 |
| | Bibliography | 264 |
| | Index | 283 |

List of Figures

| | | |
|-----|---|-----|
| 2.1 | Multidimensional Timbre Space Representation | 31 |
| 2.2 | Timbral Region Described by a Particular Model | 32 |
| 2.3 | Instrument Regions / Subspaces in Timbre Space | 33 |
| 2.4 | Example Hierarchical Structure of Timbre Subspaces Described by Axis Sets | 55 |
| 2.5 | Example Hierarchical Relationships of General Timbre Types | 55 |
| 2.6 | Composite and Prototypical Splits in a Hierarchical Structure | 56 |
| 3.1 | Instructions for Perceptual Study | 104 |
| 3.2 | Top Parts of Response Forms for Perceptual Study | 105 |
| 3.3 | Distribution of Responses Over All Sounds for All Participants with the String Type Test | 110 |
| 3.4 | Distribution of Responses Over All Sounds for All Participants with the Woodwind Type Test | 110 |

| | | |
|------|---|-----|
| 3.5 | Distribution of Responses Over All Sounds for All Participants with the Brass Type Test | 111 |
| 3.6 | Distribution of Responses Over All Sounds for All Participants with the Hammered Tonal Type Test | 111 |
| 3.7 | Distribution of Responses Over All Sounds for All Participants with the Percussive Type Test | 112 |
| 3.8 | Distribution of Responses Over All Sounds for All Participants with the Synthetic/Test-Tone Type Test | 112 |
| 3.9 | Varimax-Rotated Variable Loadings for PCA of Subjects' Test Results | 127 |
| 3.10 | Large Dot Plots of Factor Scores for Individual Stimuli Resulting from PCA of Subjects' Responses | 133 |
| 3.11 | Labelled Plot of Factor Scores for Individual Stimuli Resulting from PCA of the Responses for Subject 1 (Musician) | 135 |
| 3.12 | Labelled Plot of Factor Scores for Individual Stimuli Resulting from PCA of the Responses for Subject 6 (Musician) | 136 |
| 3.13 | Labelled Plot of Factor Scores for Individual Stimuli Resulting from PCA of the Responses for Subject 9 (Non-Musician) | 137 |
| 3.14 | Labelled Plot of Factor Scores for Individual Stimuli Resulting from PCA of the Responses for Subject 12 (Non-Musician) | 138 |
| 4.1 | The Human Peripheral Hearing System | 148 |
| 4.2 | 512 point FFT Spectrogram Underneath Time Domain Form of Stimulus 8 (martele violin) | 154 |
| 4.3 | Example Waveform with Marked Periodicity/Stability Measurement Points . | 158 |
| 4.4 | Part 1 Stability Analysis Screen Dump; Stimulus 145 | 160 |
| 4.5 | Part 1 Stability Analysis Screen Dump; Stimulus 136 | 160 |

4.6 Part 1 Stability Analysis Screen Dump; Stimulus 8 161

4.7 Part 1 Stability Analysis Screen Dump; Stimulus 25 162

4.8 Part 1 Stability Analysis Screen Dump; Stimulus 29 163

5.1 Large Dot Plots for AXESDIST Results for Equation 5.6 210

5.2 Values of Minimisation Ratio for Equation 5.8 Against Number of Dimensions;
Full Feature Set 219

5.3 Values of Minimisation Ratio for Equation 5.9 Against Number of Dimensions;
Full Feature Set 221

5.4 Values of Minimisation Ratio for Equation 5.8 Against Number of Dimensions;
First Limited Feature Set 222

5.5 Values of Minimisation Ratio for Equation 5.8 Against Number of Dimensions;
Second Limited Feature Set 223

List of Tables

| | | |
|-----|--|-----|
| 3.1 | Sounds Used as Perceptual Test Templates | 109 |
| 3.2 | Basic Statistics of Percentage Responses for Perceptual Tests Among the Participants | 115 |
| 3.3 | Pearson Correlation Coefficients of Perceptual Study Test Results | 116 |
| 3.4 | Stimulus Codes Corresponding to Particular Modal Responses for all Partici- pants | 122 |
| 3.5 | Stimulus Groupings and Colour Coding used in Subsection 3.5.5 | 132 |
| 4.1 | Balanced Main Parameters for Part 3 of the Analysis | 166 |
| 5.1 | Summary of Extracted Feature Set | 195 |
| 5.2 | Time-Varying Characteristic Metric Codes | 199 |
| 5.3 | Perceptual Test 2D PCA (First Factor) versus Features: Highlights of Strongest Correlations | 205 |

5.4 Perceptual Test 2D PCA (First Factor) versus Features: Highlights of Lesser Correlations 206

5.5 Groups Used in Subsection 5.6.2 for AXESDIST Analyses 207

5.6 Results for AXESDIST Minimisation of Equation 5.6 208

A.1 Descriptions of Sound Samples in Use 245

Acknowledgements

Tremendous thanks are due to the following people:

- My wife, Emma, for her ever-faithful support.
- My family; especially my father, mother and brother.
- Andy Tyrrell and David Howard for their invaluable supervisory skills.
- Ross Kirk and Francis Rumsey for being my examiners.
- Andy Hunt for fruitful discussions on life, the universe and everything musical.
- Terry Edhouse and the computing support team for dealing with my regular questions and problems.
- John Szymanski for mathematical supervision.
- Tim Anderson, James Angus, Tim Brookes, Darren Buttle, Richard Canham, Mike Evans, Mike Freeman, Pete French, Neil Garner, Paul Garner, Ian Gibson, David Hedley, Damian Murphy, Paul Murrin, Tony Price, Mark Pearson, Roger Peppé, Cesar Ortega, Barry Scowen, Steve Smith, Tony Tew and John Tuffen for advice, comments, and help with perceptual testing.

Financial Support

The research upon which this thesis is based was supported by an EPSRC grant (award number 94315368).

Equipment and Software Used in This Research

1. All the acoustical analysis and feature extraction software was written using Borland Pascal version 7 under Microsoft MS-DOS version 6.
2. 3D visualisation software was written with MATLAB version 4. Other graphs were produced by SPSS version 6.1.
3. Statistical analysis was performed with SPSS version 6.1, Microsoft Excel version 5.0 and also with software written by the author in Pascal.
4. Andromeda Grafix version 4.3 was used for MS-DOS screen capture. R.S.Horne's Spectrogram version 2.3 was used to produce Figure 4.2. Images were edited with JASC Paint Shop Pro version 3.11.
5. SPSS, MATLAB, Excel, Spectrogram and Paint Shop Pro ran under Microsoft Windows version 3.1.
6. For data computation and analysis an IBM-PC compatible with Intel Pentium 150MHz processor and 32Mb of RAM was used. Also, a number of Intel i486-based 100MHz/16Mb machines were used to supplement the computing power during the computationally-intensive analysis stage.
7. The audio hardware used was an Orchid NuSound card in the IBM-PC compatible, and Sennheiser HD435 headphones.
8. This document was prepared with $\text{\LaTeX} 2_{\epsilon}$ (\TeX version 3.14159, C version 6.1) using a IBM P120+ based IBM PC compatible with 16Mb of RAM running Linux (kernel version 2.0.30). Hard copy versions were produced on the following printers; HP DeskJet 870CXi, HP LaserJet 5MXSi and HP ColorLaserJet 5M.

All trademarks in this text are the property of their respective owners.

Declaration

This thesis is entirely my own work and all contributions from outside sources, through direct contact or publications, have been explicitly attributed.

© David P. Creasey

March 1998

CHAPTER 1

Introduction

1.1 Background to Research in Timbre

At the present time, understanding of the human hearing system as a whole is still largely incomplete. Mechanisms associated with the auditory path, and acoustical properties which appear to relate to important aspects of aural sensation, have been investigated in the past. Yet, there is less evidence of a universal model of hearing processes developing than is present in vision research ([15]). The modelling problem extends from the highest level of sonic entity identification, through intermediate mechanisms of sound form feature analysis, to the low levels of raw acoustical data and its treatment at the periphery of hearing. Of these, the lowest levels are best understood. This understanding was originally achieved through dissection and stimulation of the biological component parts. The majority of auditory analysis, however, occurs in the auditory cortex, which cannot be

dealt with successfully in such a manner.

A major part of the modelling problem is that auditory perception is not based on a simple set of axes. It is governed by the complex interaction of acoustical properties. The basic elements of sound are generally considered to be loudness, pitch (if a pitch percept exists), duration, and “other qualities”. The first three can usually be measured, compared, and modified with accuracy. Yet a sound cannot be completely described without the fourth major element. “Sound qualities” also represents those elements of sound which are least well understood. The complexity of this category is reflected in the lack of a clear definition, and lack of consensus as to what is implied by the fourth major element; even that it is a single major category.

Loosely speaking, “sound qualities” are known as timbral characteristics. As is discussed in Chapter 2, however, the meaning of timbre is better described by stating a “timbre space”, or region of consideration of sound qualities within the bounds of a particular sound model in particular research. The paucity of facts concerning sound qualities has led to a considerable number of interpretations of the problem. The author believes this can only be solved by appreciation of the limitations of any one particular study (and thus, definition of timbre through the timbre space of interest). Such an appreciation would prevent authors attempting sweeping conclusions from a low base of facts and the general confusion of information which pervades the literature.

This research both considers the common ground of the information from the literature, but also builds a platform of results within the confines of a particular timbre space which may be examined by future researchers. Any conclusions which are drawn are not intended to “solve” the problems of timbral form nor permit gross generalisations. What it does provide is an appreciation of previous research, the experimentation and analysis conducted as part of the present work, and information to facilitate future development.

1.2 Motivation and Aims

Understanding sound timbre is more than just an academic exercise. Hearing is a vital sense in humans, yet research into its form and function has been very limited compared to the advances in visual analysis, modification and synthesis. The structure of perception of

sound qualities and its links to the acoustical world are of great relevance to a diverse range of areas such as the following:

1. The medical treatment of auditory system defects.
2. The understanding of perceptual processes and the structure of cortical processing.
3. The relationships between natural events and the associated reactions in animals.
4. The understanding of the effects of physical characteristics of sound sources.
5. Speech recognition and synthesis.
6. Music technology in general, and the processing and creation of sound.
7. The natural development of instruments and musical form through history.
8. The development of interactive and immersive computer technology.

This thesis details an investigation of timbral form which pulls together the diverse work of other authors, and expands understanding of the form of timbre space; both structure and magnitude of effects. The most important characteristic of the work is that it considers both perceptual understanding and acoustical form, and considers the links between them. The thesis covers the following aspects of research undertaken by the author:

1. A substantial literature survey which brings together many perspectives on definitions of timbre, structure concepts, methods of timbre study, and previous results concerning acoustical/spectral forms and timbre perception (Chapter 2).
2. A perceptual study which achieves progress in basic understanding of the mechanisms of timbre perception and the mental relationships between different types of sound qualities within the data set (Chapter 3).
3. A spectral analysis-synthesis scheme which investigates adaption of viewpoint based on periodicity metrics, to achieve most appropriate time-frequency resolution depending upon acoustical conditions (Chapter 4).
4. A feature extraction and analysis scheme to investigate the importance of different parts of the spectral form in distinguishing between sound types, and thus the links

between perception and acoustical form. Also, the methods of extraction and statistical analysis, hierarchical structuring and the dimensionality of timbre space are considered (Chapter 5).

1.3 Details of Hypothesis

Timbre space is a perceptually-structured complex multi-dimensional form, which is ineffectively and imprecisely characterised by a low number of exclusive single dimensions or general descriptions, but can be represented by grouped features derived from time and frequency domain representations.

A “timbre space” is a multidimensional representation which describes the perceived and/or acoustical relationships between sonic entities through differences in their sound qualities. A timbre space region is an area of consideration in a particular study of sound qualities. Both concepts are discussed in more detail in Chapter 2.

That timbre space is “perceptually-structured” means that there are consistent logical relationships in the timbre space which describe the perception of the timbre of sonic stimuli which are contained in it. The nature of the relationships can be established through experimentation. The structure involved is a natural part of perception and has a continuous and cohesive form, as described in Chapter 3.

That timbre space is “complex” and “multi-dimensional” results from the vast range of perceptible sonic nuances and the complicated nature of auditory perception. The timbre of sounds cannot be described as simply as can the pitch, loudness and duration. Timbral perception demonstrates significant cross-coupling between those dimensions which are presently understood, apparently non-linear facets, and such difficult concepts as context variability of effect.

That timbre space is “ineffectively and imprecisely characterised by a low number of exclusive single dimensions or general descriptions” means that it is not sufficient to expect a three- or four-dimensional model of timbre to explain all the nuances of perceived sound quality (Chapter 5). Neither should it be expected that those axes will be uncoupled. Necessarily this makes accurate *general* description of the constituent parts of timbre

almost impossible in a concise manner (Chapter 2).

That timbre space “can be represented by grouped features” refers to the process of describing timbral change through combinations of primitive features. The primitive features are simple acoustical elements which relate to known perceptually important features of the sound model (the time-varying frequency spectrum in this research). Those features are combined in different ways to describe differences between stimuli in a particular timbre space structure (which is a hierarchical form in this research, Chapter 5).

That those groups can be “derived from time and frequency domain representations” means that timbre perception can be related to features of the time-varying spectral form. The links between timbre perception and the acoustical nature of sounds can then be established (Chapter 5).

1.4 Contributions to Knowledge Made by This Research

This thesis details the most comprehensive study of sound timbre as a whole to date. It pulls together a large range of related aspects to provide a balanced analysis of the topic. It describes previous research, discusses definitions and structures in depth, contains studies of both the perceptual and acoustical aspects of timbre, and provides pointers for future research work. It uses the largest range of sounds and timbral features of any study to date.

Chapter 2 is concerned with previous research in timbre, the different structural forms associated with it, and different definitions concerning timbre and timbre space representations. It is the most comprehensive overview of the sound timbre literature and concepts to date. In particular, the in-depth consideration of timbral definition, factors affecting timbre space, dimensionality and structure are novel. The Chapter contains a review of those time-varying frequency spectrum aspects previously found to relate to timbre perception.

Chapter 3 is concerned with a study of perceptual similarity judgements between the input stimuli, a statistical analysis of the result structure, and the implications for understanding of the structure of timbral audition. The study considers the perceived relationships within a stimulus set of 153 sounds relative to templates of 4 stimuli for 6 timbre families. The

conclusions are wide-ranging and indicate that timbre is a natural part of perception, timbral relationships are perceived in a logical, structured, continuous and overlapping manner. The study is novel in its scope and range of conclusions.

Chapter 4 is concerned with analysis and synthesis using a time-varying frequency spectrum model, with adaptive viewpoint properties to achieve appropriate time-frequency resolution. The system works with a wide range of sounds without user intervention but has adaptive time-frequency trade-off linked to the periodic/stability conditions within the stimuli without a knowledge-based system and with a frequency-domain result. The implementation of this particular combination of features is novel.

Chapter 5 is concerned with the extraction of 335 acoustical features from the spectral forms produced by the method of Chapter 4, and a statistical consideration to find those features which describe perceptual differences in timbral form between the 153 stimuli. The experiments are again more wide-ranging than previous research, and use statistical techniques which examine the relationship between perceptual and acoustical form in an objective manner. The study considers hierarchical decomposition and the dimensionality of timbre space in an empirical manner, which have not been done in such a comprehensive way before.

A more detailed consideration of the knowledge provided by this thesis is given in the Conclusions.

1.5 Ambiguous Terms

These terms have common multiple meanings in the literature and are important to understanding the following text. Other, less ambiguous, terms are explained in the Glossary.

1. **Analysis.** This refers to both the process of transforming sound data into a manipulable/distributed format (such as a Fourier transform), and also examining the detail and structure of information.
2. **Harmonics.** These are low-bandwidth spectral elements at approximately integer multiples of a fundamental frequency, integer numbered sequentially from 1 for the

fundamental.

3. **Overtones.** These are harmonics above the fundamental frequency and are numbered as the harmonic minus one. Most, but not all, authors adhere to this form.
4. **Partials.** These are low-bandwidth spectral elements (components) without frequency constraint. Some authors mean harmonics when they refer to partials.
5. **Sample.** This, in general, refers to an instance of sound, rather than a time-quantised time-domain value used to represent the acoustic wave at a particular point in time. The latter meaning will be used when discussing analysis from, and resynthesis to, the time-domain form.
6. **Spectrum.** This will be used to refer to time-varying frequency spectra in the audible frequency range (that is, the spectrogram domain), unless otherwise stated.
7. **Synthetic.** This term is used to refer to sounds which have a form which is not closely related to a transformed version of an acoustic wave from a physical source. It is an imprecise concept, and should not be taken to mean either that a sound instance is a simple test tone or cannot sound as if it is from a “natural” source, but rather that it has been “constructed” by a synthesis technique, rather than simply synthesised from an analysed physical source. The perceptual meaning of synthetic is expounded further in Chapter 3.

CHAPTER 2

Perspectives and Research into Timbre

“... timbre is crucial to the understanding of the sonic world.” [107]

“Ideally we require a way to “measure” or order these degrees of difference allowing us to articulate the space of sound possibilities in a structured and meaningful way.” [226]

2.1 Introduction

This chapter presents a wide-ranging overview of the nature of sound timbre. Definitions, associated structural forms and previous research work are all considered. The information

presented here is used as a basis for the exploration of timbre in Chapter 3 (the perceptual aspects) and Chapter 5 (how acoustical aspects are linked to perceptual forms). This chapter discusses:

1. What timbre and timbral spaces are.
2. What viewpoint this research takes on timbre.
3. Previous relevant studies of timbre.
4. The relationships between high and low-level descriptions of timbre.

2.2 Definitions of Timbre : Overview

“timbre (tēbr, 'tæmbə(r)), *sb.*³ [a.mod.F. *timbre*: see TIMBRE *sb.*¹ and ². From the sense ‘bell’, ‘small bell’ (see TIMBRE *sb.*²) arose that of ‘sound of a bell’, ‘sonorous quality of any instrument or of a voice’, and finally that of ‘character or quality of sound’ (= Ger. *klangfarbe*), in which the word has passed into English use, retaining its French pronunciation.]” [148], p.100

This section outlines the broad range of views in existence, concerning the definition and character of timbre. The concepts involved are not always referred to by the same word. Sometimes names such as tone colour, tone quality, sound colour, sound quality, *Klangfarbe* and timber (sic) are used. Attempts to pin down exactly what timbre means have occurred since Helmholtz’ classic work ([83]) and possibly before. The essential problem is that timbre is often treated as a catch-all category for any dimensions of sound that have not been formally specified to date:

“Timbre tends to be the psychoacoustician’s multidimensional waste-basket category” [122]

This approach is typified by the most quoted timbre definition; that of the American National Standards Institute (ANSI):

“Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar.

Note: Timbre depends primarily upon the spectrum of the stimulus, but it also depends upon the waveform, the sound pressure, the frequency location of the spectrum, and the temporal characteristics of the stimulus.” [1]

This attempts to bring together a large number of viewpoints, which is possibly why it still provokes considerable discussion in the literature. The first part clearly states that pitch and loudness are not part of timbre. “Similar presentation” is broader. The consensus would appear to be that this implies that environment, or room acoustics, reproduction effects and spatial location are not part of timbre ([8], [71]). However, this necessarily excludes environment-like timbral effects, such as echo and reverberation. For example, very heavy reverberation, or distortion through overdriven reproduction equipment can totally change the essential “tone colour” of a sound. If these are excluded from timbre, then a new dimension of sound “environment effects” must be created.

The second part of the ANSI definition attempts to qualify the previous strong statements, seemingly to make it a less negative description:

1. Saying that timbre is dependent primarily on the spectrum recognises the “spectral view” of timbre, but fails to state whether this is the overall effect (long-term average spectrum) or a continuously variable spectral form with time (as in this research).
2. That timbre “also depends” upon the waveform is confusing, in that a waveform can completely describe a sound, without recourse to a spectral view. Again, this possibly means an “average” waveshape. Furthermore, waveforms can appear radically different, through differences in phase of frequency components, yet sound the same or very similar ([159]).
3. Including sound pressure in the note rather invalidates the loudness part of the original definition.
4. Adding “frequency location of the spectrum” identifies the importance of “brightness” in timbre through the centroid of the spectrum. This is considered later in the chapter.

5. Finally, “temporal characteristics of the stimulus” indicates the important role of the time- and frequency-varying characteristics of the spectral form, such as the amplitude curve and movement of frequency partials. This is of considerable note, particularly if the earlier parts of the definition propose a static view of the waveform/spectrum. This will also be covered in more depth later.

A particular failing of the definition is its emphasis on the spectrum/waveform, as these imply a *particular instance* of sound. For example, two similarly (i.e. as identically as possible) played similarly pitched notes will never be identical in terms of their spectrum/waveform ([80], [137], [226]). A strict interpretation of the ANSI definition must judge them to have different timbres. But, for all practical purposes, the timbres may be considered identical by listeners ([172]). Therefore, it is important to consider carefully what level of detail and at what scale the examination of timbre occurs. In fact, the answer depends upon the situation. Furthermore, it is unwise to restrict timbre to a time-varying spectral definition, when timbre can equally be described in terms of other analysis-synthesis models, such as physical (which might describe timbre in terms of resonances and drivers) or granular (where timbre might be considered in terms of density and types of grains of sound) for example.

The ANSI definition is the most famous of timbre descriptions and represents the middle ground. However, many other authors have postulated alternative viewpoints.

“**timbre** (*Acous.*). The characteristic tone or quality of a sound ” [28]

“By timbre is meant the distinguishing or characteristic quality of a sound; it is by their timbre that we recognise an instrument . . . regardless of the pitch or intensity of the note it is sounding.” [96]

These are more basic definitions of timbre. They highlight characteristic qualities as being important rather than the mechanism by which such qualities are apparent (such as the spectrum or waveform). That is, it is *perception* that is vital, not low level descriptions. Some authors limit their definition further:

“Sound color is a property or attribute of auditory sensation; it is not an acoustic property . . . sound color pertains to the steady-state portions of

sounds but not, in general, to their beginnings or endings.” [190]

“[Timbre] considered in its most narrow sense . . . [is] that attribute of sensation in terms of which a listener can judge that two steady complex tones having the same loudness, pitch and duration are dissimilar.” [157]

Here the emphasis is placed on the “steady” nature of a tone. That is, limiting timbre for purposes of a particular investigation, to the nominally intransient part of the sound. Plomp ([157]) also adds duration into the excluded categories. Others have taken a different view, either defining timbre as an instantaneous spectral composition separate from the changing characteristics over time ([180]), or encompassing time completely:

“Timbre is defined as the attribution of spectromorphological identity . . . [:] a field of interactive behaviour, played out in the relations between spectral space and temporal change.” [194]

“First of all, we can decide very well whether a sound is tonal or noiselike . . . Secondly we have, roughly speaking, spectral characteristics . . . Thirdly, there is the time envelope . . . Fourthly, there can be changes within the signal . . . micro-intonation. Finally, there is . . . the prefix.” Schouten in [156]

Schouten takes a more complex view of timbre, breaking the representation down into non-orthogonal feature sets in a more structured way than the ANSI note. Recognising that the overall model of analysis can be viewed from different perspectives is important. Saying that a sound can be represented by a time-varying spectrum is not as useful as stating those parts of the spectral form which have particular significance, even though they overlap and are fully represented by the overall model. Thus the quote above breaks up the spectrum into components such as noise/non-noise, “spectral characteristics” (by which it is assumed that aspects such as centroid, overall gradient, harmonic composition and so on are implied), amplitude envelope, micro-intonation (movement of partials within the spectral form) and the attack/onset portion. These will all be discussed in greater detail later in the chapter.

“[Timbre] ... does not exist in the ‘real’ World as an object. It is an attribute of musical tone that is abstracted from the entity that we call ‘a musical tone’ ... Timbre is not even the only attribute of tone connected to tone quality: consider the ‘density’ and ‘volume’ of S.S.Stevens ([198]). ... there cannot be a fully satisfactory single operational definition of something that is so abstract, multi-faceted, and elusive.” [75]

“Timbre is as much an intentional process, an ideology, an attitude, as a physical characteristic.” [213]

These quotes take a much more radical view of timbre - that timbral definitions are attempting to capture something that is at a much higher level than that at which it is normally interpreted. It indicates that timbral form is fundamentally a perceptual complex. Furthermore, [75] makes a distinction between tone quality (as an overall attribute of sound) and timbre (as a particular subset). It is apparent that:

“[Timbre has] extraordinary intricacy, difficult determinability and conceptual expressibility” [130]

Lastly, it is seen from considering different quotes, that there is further confusion as to whether timbre only distinguishes between instruments, or whether it also refers to the nuances of tone quality characteristics of individual instruments.

2.3 Definitions of Timbre : Specific Characteristics

The previous section gives a general overview of what timbre means to different authors, but it is desirable to be more specific, as follows:

2.3.1 Timbre as a Distinguishing Quality

The simplest way of describing timbre in perceptual terms is as a sonic character distinguishing quality. In terms of instruments, this may be considered on two levels:

1. As a means by which it is possible to perceive one instrument as different from another (distinguishing on a macro scale).
2. As a means by which it is possible to perceive nuances of sound quality; differences resulting from playing style, differences between sounds from different examples of the same instrument and so on (micro scale).

However, to limit the consideration to traditional instrument sources is to miss the point somewhat ([102]). In this research, timbre has been considered to be that quality which distinguishes any sound quality/colour, either between sound sources *or* within the sound source. Such a viewpoint on the distinguishing quality of timbre has been explicitly stated as far back as 1954 ([68]), but has not always been followed in the literature; many authors restricting the distinguishing principle to the first level only.

2.3.2 Timbre as an Absence of Other Quantities

Timbre is traditionally defined as those aspects of sound which are not the following quantities:

1. Pitch. (for example [1], [13], [15], [83], [112])
2. Loudness. (for example [1], [13], [15], [77], [83], [112])
3. Duration. (for example [13], [71], [149], [157])
4. Presentation/environment. (for example [1])

However, in the modern context it is apparent that timbre is not *independent* of those quantities, but that (depending upon the timbral definition in use by a particular author) they are, at a significant level, un-timbral. Pitch, loudness and duration allow relatively unambiguous ordering along a single scale, acoustic and verbal description, and direct measurement using clearly defined parameters, with which presentation/environment and timbre fail to comply ([201]). Part of the aim of timbre research is to find the currently missing descriptors, parameters and scales of relevance.

2.3.3 Timbre as a Non-Independent Set of Tonal Attributes

Treating timbre as an independent area of sound is to ignore the known perceptual and acoustical couplings between the various parts of sound (pitch, loudness, duration, environment, timbre, spatial location) established in previous research (for example [59], [71], [92], [102], [129], [151], [185]). This is more than an interesting argument, in that models used to investigate timbre necessarily link those aspects together. This means that the investigator must define at what point the consideration ignores or includes such aspects in the investigation of sound quality. From a practical perspective, it might be considered desirable to equalise sounds in the “non-timbral” dimensions to make comparisons easier ([71]). However, the couplings are strong enough to cause equalised sounds to be significantly unrepresentative of the original instrument in many cases (for example, changing a short high pitch sound, and a long very noisy unpitched sound, both to moderate lengths and pitches), and fails to aid in understanding natural differences between them ([102]).

Environmental conditions at a low level might be considered “reproduction characteristics”. At some point, reproduction characteristics might reasonably be considered timbral modification through environment-like effects, such as reverberation, distortion, and flanging/phasing. As regards pitch, the most obvious consideration is brightness of sound resulting from the fundamental frequency. A set of harmonics with a fundamental at high frequency will sound brighter than one at lower frequencies. Additionally, some sounds have no pitch associated with them, or a very ambiguous pitch percept which is difficult to quantify, let alone equalise out of the model ([130], [193]). Changing the loudness of a sound necessarily changes the perceived frequency balance, due to the different frequency response of the ear at different intensities. Also, loudness/volume can be implied by timbral cues, not simply intensity ([199], [215]). Finally, very short duration sounds have a very different tone colour to those which are longer ([77]). Yet, grains of sound can have pitch, spectral contour, onset characteristics and so on ([226]). It is necessary to consider how much these and other effects can be detached from consideration.

2.3.4 Timbre as a Multidimensional Entity

To say that timbre is a single attribute of sound is incorrect. For considerable time ([71], [77], [113], [217] and others) it has been apparent that timbre is neither a singular axis of sound, nor one whose characteristic components can be described more than partially along a well defined scale such as Hertz or decibels. As will be shown later, at the present time, such axes as have been identified are neither comparable (that is, different axes can have very different magnitude of aural effect), nor orthogonal (that is, they are, generally, perceptually coupled, [120], [194], [217]).

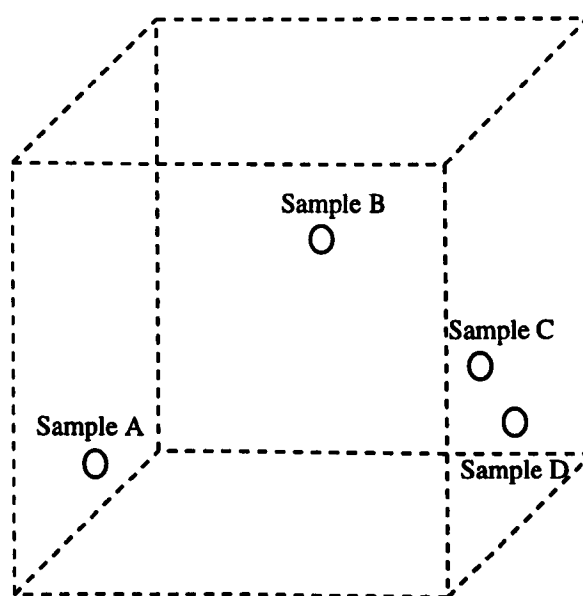


Figure 2.1: Multidimensional Timbre Space Representation

A useful way of visualising what timbre means, is the “timbre space” representation. This is a multidimensional space where distance represents timbral dissimilarity between sonic entities. Figure 2.1 shows a three dimensional example of the relationship between individual sound samples. The dimensions of this construct enable classification and navigation in the space. If the fundamental orthogonal dimensions of timbre were known, these would form an hypercube describing every audible timbre type (the “overall” timbre space). Grey ([71]) suggests that timbre space may be considered a combination of both low level (acoustical) properties and higher level cognitive differentiation between stimuli.

When a version of timbre is defined it implies a *region* within the overall timbre space, with dimensions corresponding to the proposed analysis-modification-synthesis model

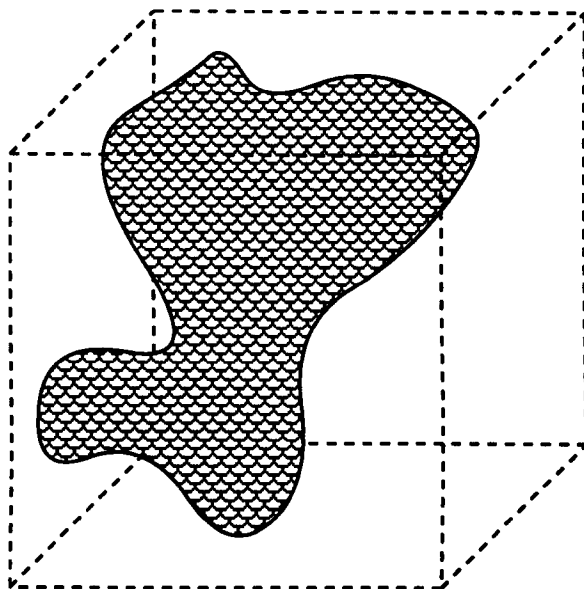


Figure 2.2: Timbral Region Described by a Particular Model

(Figure 2.2). As has been shown by previous research, the probability of the model's axes corresponding to the fundamental orthogonal axes of timbre space is extremely low.

As timbre space is a navigable construct, it is possible to consider the way that sounds are related to each other within the multidimensional space (Figure 2.3, [123]). Thus timbre space represents the nuances of sound, as epitomised by the differences between samples from one sound source (performance characteristics). One instrument can produce a considerable range of timbres, yet they are all perceived as a coherent set ([82]). Larger differences in timbre space are equivalent to the differences between instruments, although in the modern context, both small and large timbral change can be produced as part of the same acoustic source, particularly in electroacoustic music. Timbre space is not necessarily perceived as a plain multidimensional set of axes, but might be an hierarchical series of axes or augmented by specific attributes. These aspects are discussed in Section 2.7.

Combining an instruments' timbre space with its other attributes (pitch, loudness and so on) creates an "instrument space" form in the sound model ([211]).

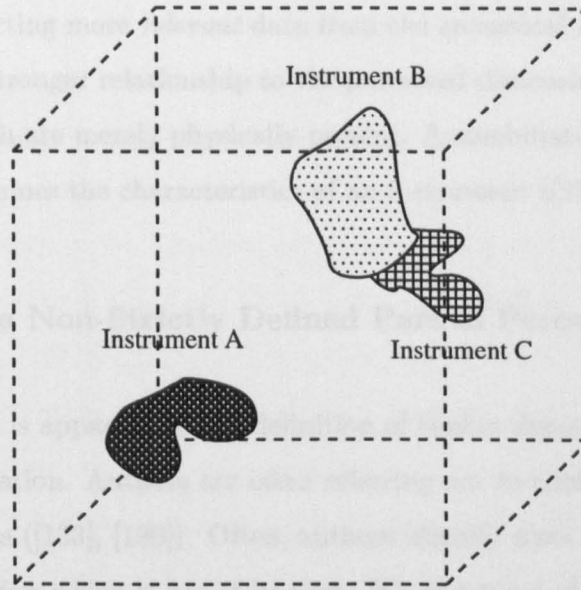


Figure 2.3: Instrument Regions / Subspaces in Timbre Space

2.3.5 Timbre as a Complex Composite

Timbre is a composite of a number of related perspectives produced from the model used for investigation. Such a model can be viewed in different ways such as:

1. A set of instantaneous model conditions (for this research, frequency spectra) linked together by time.
2. Large-scale abstractions from the model (such as amplitude envelope over time, or spectral gradient and centroid).
3. Internal structure of the model (harmonic relationships, attack to average amplitude ratio, and so on).

Which viewpoint is most important depends upon the type of effect that is being looked for. For example, in distinguishing a limited range of orchestral instruments, Grey ([71]) found that it would seem to depend upon the uniqueness of the spectrum as to whether the spectrum or its evolution is most significant. Similarly, different features abstracted from the model representation will have different prominence in different situations.

Timbral features (axes) may be extracted from these different viewpoints, to produce a set of parameters which may be used to interrogate, or used in reverse to manipulate, timbre.

This process is abstracting more *relevant* data from the acoustical dimensions of the model, which bare a stronger relationship to the perceived dimensions of timbre, as opposed to those which are merely physically present. A combination of these acoustic variables, then, determines the characteristics of an instrument ([22]).

2.3.6 Timbre as a Non-Strictly Defined Part of Perception

From the literature, it is apparent that a definition of timbre depends upon the *type* of timbre under consideration. Authors are often referring not to timbre in its entirety, but a range of timbral effects ([133], [190]). Often, authors identify a set of timbral exclusions which indicate the region which is *not* of interest. When a piece of research acts “on timbre”, it actually acts on a limited timbre space region defined by the following quantities:

1. Exclusions. That is, specific sonic attributes with which the system is not concerned. Classically, this means pitch, loudness, duration and environment, but there are other general classifications which do not relate directly to the sort of sounds under consideration (see next point). That is, limitations such as the range of frequencies under concern, or excluding certain sorts of model considerations, such as formant structures.
2. The sounds used with the system. These may be limited to simple synthesised complexes ([185]), a subset of instrument sounds ([71]), vowel sounds ([190]), sonar sounds ([196]), or broader such as orchestral music ([30]) or animal sounds. In particular, it has been suggested that speech and musical sounds are processed in different parts of the auditory cortex, and thus the timbre may be analysed in different ways ([80], [139]). The input dataset determines the solution space ([157]), limits universal applicability ([13]), and limits the number of important dimensions of timbre that are likely to be uncovered ([15]). In particular, the use of appropriately generated synthetic sounds aids in the examination of the effects of individual characteristics, without the acoustical and/or perceptual clutter of other (potentially interfering) aspects. Alternatively, sounds from “natural” sources are known to have the characteristics pertinent to their perceptual interpretation, which are worthwhile investigating, within their form. This enables fine examination of

that which causes, say, a clarinet to sound as it does. Synthetic sounds can fail to enable examination of all timbrally pertinent aspects. However, the way that analysis and operations on data is performed is enhanced by limitations: For example, a system only concerned with singing can make extensive use of formant operations, and harmonics. However, one that is also concerned with waterfall sounds must place much more emphasis on the noise aspects of sounds. Furthermore, it is no longer possible to expect the frequent presence of vocal formants, or the existence of fundamental frequencies between about 87 and 784Hz (minimum range, [103]). Using particular groups of sounds can help identify important distinguishing features within those groups, which might be swamped in the consideration of a much wider-ranging set of inputs. There is also the question of how timbral experiments are to be controlled. The more simple the input data, the more easy it is to diagnose which aspects lead to particular perceptual differences, but the more likely it is that information concerning complex timbral aspects will be lost.

3. Timbral variation. The sound model's quantity of input data, and the way in which it is used and interpreted, limits the coverage of the timbral region of interest, and thus the applicability of results to different sounds ([143]). Every orchestral instrument, for example, can produce a considerable range of timbral change, which describes an instrument control subspace. A human voice produces considerably more variation in timbre than a single orchestral instrument ([80]). But, a spectral analysis of a single note is only a snapshot of the instrument, not a description of the entire system ([3]). Similarly, a physical model may describe the general structure of the instrument, but not the finer characteristics which give an instrument particular qualities. It is important to decide whether the differences *between* sound sources (macro-timbre, large distances in timbre space) and/or the timbral variation *within* a sound source (micro-timbre, variation within a smaller region of timbre space) are of concern in a particular timbral system ([71]). Further, it is often of interest to consider the commonality of timbral quality of a number of sound sources, such as "muted", "bright" or "percussive" effects. The implied distinction between "large" and "small" timbral change as a concept is becoming more blurred as time progresses, particularly in electroacoustic music ([226]). That is, it is becoming unreasonable to consider that timbral variation is only a concept of distinguishing between instruments *or* within instruments' timbral subspaces.

4. The ability of the implemented model to cover the entire region of interest. For example, a single Fourier analysis has a trade-off between frequency and time information which leads to a particular perspective of the timbral form of a sound. Furthermore, if parameters are derived from that analysis, it imposes a further distinctive interpretation of the information. If there are then limitations on how it is possible to manipulate that information, then the ability to navigate the timbre space is curtailed. The model may also limit ability to (re-)synthesise with high accuracy.
5. Context. Whether sounds are considered in isolation, or in simultaneously sounding groups, or in succession with transitions between them, significantly affects the nature of the timbre space under consideration ([20]).

2.3.7 Timbre as a Highly Coupled Complex

While fundamental orthogonal dimensions of timbre (or indeed, sound as a whole) elude capture, consideration must be concerned with how varying one timbral parameter will affect others. It seems that it is the nature of timbre that nothing is entirely independent. Altering one characteristic of the sound model form for several source items may achieve a particular effect, yet when the effect is applied to a different item, it can change other aspects of the timbral form as well. On the other hand, the source materials' structure can impose a *lack of* ability for a feature to alter the desired timbral parameter. For example, if brightness is recognised as a property of increasing spectral centroid, then it is impossible to alter the brightness of a low frequency pure tone without changing the pitch or fundamental structure of the tone (which is unacceptable). As is apparent from this example, cross-coupling is a larger concern in systems that manipulate timbre, compared to those which are only analysing timbre. Analysis is often hard, but perceptually correct manipulation is harder. It remains to be seen if it is possible to find new orthogonal axes from those highly coupled ones identified so far (principal axes/factors) which have a meaningful and universal application. Using a top-down approach, it may be possible to dissect the overall representation by considering statistical methods to reduce the universal data space to orthogonal factors, leaving a problem of complex acoustical interpretation. From the bottom-up, it is necessary to know all the low-level parameters appropriate to timbre in order to attempt to find all common attributes and uncouple the dimensions.

2.3.8 Timbre as a Non-Linear Construct

Despite auditory information generally behaving in predictable ways ([80]), timbral forms do not generally appear to display “simple” characteristics overall. For example, when morphing between sounds, linear steps in the acoustical representation rarely equate to linear perceptual steps. Similarly, due to the cross-coupling considered in the previous subsection, and the dependence on the structure of the sound, it is unreasonable to consider that a timbral parameter can have equal magnitude or type of effect in modification for all sounds. Furthermore, masking effects between timbral aspects play a role, which cannot be ignored in some systems. That is, a timbral effect on its own may have a prominent psychoacoustic effect, yet in the presence of another effect may no longer be perceptually relevant.

2.3.9 Timbre as a Continuous Quality

Modern timbre research is based implicitly on the idea that sounds do not represent discrete, isolated data points. That is, timbre space is a continuum. Because of the non-linear nature of sound qualities ([46]), perceived transitions between timbral “states” can be very abrupt, but the weight of evidence suggests a non-categorical system ([71]). As shown in Chapter 3, the differences between instrument tones is not always obvious, and confusions of source can occur quite readily. By considering acoustical parameters, it is also found that there are no large gaps between categories of instruments at a low level (Chapter 5). Recently, the tools and processing power has existed to investigate the perceptual ambiguities present in sonic interpolations ([209], [226]). It is often assumed that any theories of timbral change or modification represent scalable vectors ([12], [71]) in timbre space which, at least to some extent, can be translated to other parts of timbre space and maintain their perceptual effect, and/or combined with other effects. Such forms allow the development of “timbral object” representations ([20], [35]). These assume a continuum of timbre to exploit, not a loose collection of data points.

2.4 Definitions for this Research

“... the scientist finds it thin to try to nourish his understanding by the ingestion of semantic disputation.” [198]

In the final analysis, it is possible to argue forever about what the one term “timbre” actually means. This is especially so because it is so hard to specify what it is, rather than that which it is generally accepted not to be. A perceptual definition of timbre can be as broad as *timbre is the character of the sound*, or *timbre is those aspects of a sound which identify it to the human aural system as being unique*. However, it is operational definitions which are of importance, as follows:

Definition 1

Timbre space is the multidimensional system which describes disparity in sound qualities through geometric distance.

Definition 2

A timbre space region is an operational hyperspace described by a combination of sonic character attributes, which parameterise the audible multidimensional space of interest.

From this, it is apparent that the implicit aim is to find sonic character attributes which are important in parameterising particular timbre space regions, which can distinguish between sound quality types of interest

Definition 3

A sonic character attribute (SCA) is a parameter of the timbral model derived from an acoustical representation of sound, which if progressively modified would be perceived as a progressive change to auditory sensation, which is, at least in part, to sound qualities other than pitch, loudness and duration, and is not a parameter of spatial location.

These definitions state that the region of interest is that defined by the attributes of control. The definition of SCAs deliberately states that:

1. They are derived from an analysis model. That is SCAs are not, fundamentally, perceptual parameters, they are measurable quantities which are linked to timbral attributes.
2. They produce progressive change. That is, SCAs are control parameters, which relate an acoustical to a (some) perceptual change(s).
3. They change sound characteristics other than pitch, loudness and duration *at least in part*. This states both that they are not concerned with pitch, loudness and duration, but also that they may affect these as a byproduct of the “sound character”-like effect. As such, even pitch, loudness and duration can be considered SCAs as they all alter other qualities.
4. They are unconcerned with spatial presentation.

It is not expected that SCAs will be independent orthogonal dimensions of sound.

Additionally, whether timbre space is static or has a time-varying aspect depends upon the sort of SCAs with which it is described. In this research, the aim is to find static quantities that describe particular aspects of timbral relationships (for example, attack rate), rather than general areas (such as the amplitude envelope). This is considered in more depth in Chapter 5. The above definitions do not attempt to characterise the entirety of timbre as envisaged by all researchers. Thus, for purposes of this research, “timbre” is an abstraction from the defined timbre space;

Definition 4

Timbral characteristics are the subjective correlates of the composite effects of SCAs in the timbre space defined in a piece of research.

It is important to note that this research deals with both single instances of sound, but is also interested in how instruments relate to one another where;

Definition 5

An instrument is a unified region of timbre space defining a coherent set of sounds which appear to come from the same source in some manner.

This definition is after [51] and [41]. It is not at all limited to “traditional” instruments (as in, sound sources used commonly in the Western music tradition), but may refer to a region of timbre space which is leaf rustling sounds, or FM-synthesised sounds which are similar to each other, for example.

Having created such a broad base, it is then necessary to reduce the consideration to a particular area of interest. This research considers a timbral region limited by:

1. Analysis by time-varying frequency spectrum model.
2. A model used for analysis with adaptable time-frequency resolution (Chapter 4).
3. Equalising reproduction effects, as far as possible.
4. Using input sounds which conform to the following specifications:
 - (a) Sound sources which are generally in common use as musical instruments, with some synthetic test signals, and excluding vocal sounds and “everyday” sounds.
 - (b) Sample lengths of ≤ 5 seconds to limit necessary processing time.
 - (c) Samples of a single isolated note, with generally static pitch, if pitched.
 - (d) Samples chosen to represent a broad range of instruments, but also variations within those instruments where possible.
 - (e) Samples which are discrete instances of sound.
 - (f) Source-focused rather than process-focused transformations ([226]). For example, with traditional instruments, sound source variations normally associated with playing style rather than by time-domain studio techniques (flanging, distortion, artificial reverberation and so forth), or inherently time-domain electroacoustic manipulation (brassage, shredding, granular techniques and so on).
5. A set of samples sufficient to cover a considerable range of instrumental sounds conforming to the above specifications which exceeds the numbers of instruments used in many previous studies (such as [71]), yet is not so large as to preclude a perceptual comparison study (see Chapter 3), or entail unreasonable computational time. This balance resulted in the use of 153 sounds in this study.

6. The extraction of spectral features which are designed to cover a broad range of forms previously identified as being significant in the literature, augmented by new forms, and which in general might be manipulable as well as analysable.
7. The consideration of isolated samples, not the study of continuation and transitions between sample elements, which is a very complex topic in its own right.

The analysis procedure is described in Chapter 4, the sound samples used in this research are described in Appendix A, and the features which are extracted are covered in Chapter 5.

2.5 Quantities and Qualities of Input Stimuli

“... the set of sounds that has been used in experiments is extraordinarily limited. No research has sampled notes across the playing range of one instrument and many studies have used unrepresentative timbres because the notes were at the extreme playing range of an instrument.” [80]

As described in Subsection 2.3.6, the types, distribution and number of input sounds are important determining factors in the type of timbre space being investigated, and thus the results an investigator achieves from experiment. Section 2.4 included a brief overview of the sort of stimuli being used in this research, but it is important to consider the appropriate choices in more detail.

As regards quantity, it would seem that, previously, research has often concentrated on relatively small numbers of sounds. For example, 9 in [157], 16 in [71], 16 in [93], 16 in [107], 17 in [142], 18 in [43], 21 in [102], 27 in [201], 36 in [105], and 40 in [87]. On the one hand, this limits the universality of any conclusions which may be drawn from the work, as the timbre space is quite limited and may bias the results of the study ([143]). However, restricting the number of stimuli has benefits in reducing computational time, and experimental time in general (such as the time to perform perceptual tests, preparing the data and so on). Additionally, a timbre space can be constructed in a particular manner, such as attempting to distribute timbral data evenly (if possible), such as a distribution around the “orchestral” space, which can help in visualising and interpreting results, as

timbres do not form heavy clusters.

A smaller number of researchers have considered a larger number of sounds, which is done in order to create more density in timbre space. This allows more effective investigation of timbral nuances, as well as the significant differences in timbre space, while still considering a broad range of timbres. For example, 98 sounds were considered in [149], and 102 in [56]. It becomes harder, however to maintain a balanced distribution in timbre space, because it is possible to obtain significant numbers of stimuli in close proximity (from the same instrument) and significant variation from some instruments, but not others. This, in general, results in a lot of points grouped together and a smaller number at intermediate distances between instruments. There is also a problem of availability of the samples recorded under similar conditions. This research has aimed for a larger rather than a limited range of samples, to investigate both the nuances and larger differences in timbre, resulting in a total of 153 sounds. It is worth pointing out, though, that this still represents a very limited set of timbral instances, despite being significantly more than most studies in the past. However, it may improve the applicability of the results and methods used to obtain those results over a larger proportion of timbre space than before.

Overall, sampling adequacy in such a complex system is very hard to define. What is apparent from studies involving smaller numbers of stimuli is that the results have been verified by other research, particularly in the case of Grey ([71]). This indicates that from smaller numbers of sounds, if appropriately distributed, it can be possible to extract major axes of timbre. Assuming an hierarchical timbre structure, it should be expected, then, that by increasing the number of stimuli considerably it is possible to find progressively more minor nuances (and associated descriptive axes) by increasing the sampling of the data space. A greater number of data points also may lead to a greater number of potential solutions when searching for particular arrangements of points and the axes that achieve those arrangements, as shown in Chapter 5.

As regards sound qualities, there is to be found a spread of sound types from test tones and other simple stimuli, to samples taken from physical instruments in the literature. There is also a considerable range of possibilities in between, such as hand-simplified representations of real instrument sounds ([71]), and FM syntheses with qualities similar to traditional instruments and hybrids ([102]). What is noticeable, however, is that the majority of timbre research has either been concerned with elemental processes of hearing

based on very simple stimuli, or traditional Western instrument types. As mentioned previously in Section 2.4, this research is also mainly concerned with Western/orchestral type instruments. However, there are also a considerable number of percussive sounds from many traditions, synthesised sounds pertaining to physical instruments' characteristics, test tones, and also a few miscellaneous sounds for control purposes.

When choosing the types of stimuli to use, researchers often use “synthetic” sounds. If such sounds are synthesised from scratch, they can have their qualities specified precisely, to enable even distribution over the timbre space of interest, or to investigate particular aspects of audition. Similarly, they may be specified with particular pitch, loudness, duration and environmental conditions, and produced in unlimited quantities. These are difficult to achieve with samples of physical instruments. However, it is of particular interest to researchers to understand the properties of traditional instruments, and their complex nature is rarely matched by synthetic reproductions. This is important when attempting to study what is going on in the auditory cortex:

“...it is by no means self-evident that the laws found by using elementary stimuli (which serve as a basis for sensation) still hold true when highly complex stimuli (which serve as a basis for perception) are used.” [139]

Traditional instruments are known to have properties of interest. They have evolved to share spectral characteristics, despite any physical necessity for this ([10]), which indicates that they have become tuned or selected to fit the properties of the ear. Synthetic sounds are liable to lack some of those properties, mainly because the processes of timbre perception are not understood, which is why instruments are being investigated in the first place. An alternative way of considering timbre is to use the sorts of sounds which make up the majority of stimuli in the World, so called “everyday” sounds ([67]) which do not fall into the neat categories that pitched instruments often do ([226]). These have not been employed in this research. The stimuli which have been used have not been extensively equalised in the “un-timbral” dimensions, as this would have been very difficult without greatly altering their natural timbres (Subsection 2.3.3).

A number of authors have used the McGill University Master Samples (MUMS) as a source of recorded instrument sounds. Papers in which MUMS were used include [87], [93], [105], [154], and [176]. Although the samples are, in general, of traditional instruments and

percussion, rather than of a broader range including other, more everyday sounds, they represent a common point of reference for different researchers and so have been used for the majority of stimuli in this research. The particulars of the stimuli used in this research are given in Appendix A.

2.6 Use and Application Concepts

Research in the area of timbre necessarily assumes a particular viewpoint for investigation. This is limited by the timbre space (Subsection 2.3.6 and Section 2.4), firstly, but the techniques used to investigate that timbre space ultimately define what knowledge is gained about the perception of timbre.

For example, the researcher may be concerned with the structure of timbre perception, so as to find a form which can convey the relationships between sonic entities in terms of degrees of difference or similarity. Such methods can be conducted unburdened by the necessity to be able to manipulate specific acoustic properties of the elements, by performing perceptual similarity tests. Alternatively, if the focus is totally on manipulation, then the necessity is to find transformation techniques which produce the desired perceptual change, without being as concerned with the actual perceptual relationships or low-level acoustical features involved, just that the change is controllable. For example, morphing between two sounds based on spectral forms can sound perceptually reasonable without the morphing operation needing to break the sounds down into particular timbral parameters such as “string-like” or “bite”. Alternatively, it is possible to focus on the model and how individual aspects contribute to the variation between sounds through analysis and/or manipulation. Different research combines different aspects of the above viewpoints.

When attempting to understand the processes of timbral perception and how they relate to the model under consideration, the level of detail becomes important. If the aim is to find *major* acoustical contributors to variance between the sounds that compose the input data set, then it is expected that the system will have a low number of dimensions. Potentially, such a form could have $N - 1$ strong axes and a residue axis to account for the remaining variance. Or, the residue components may be distributed among all axes. A different approach is to consider that timbre is an hierarchy of axes, where it is possible to

find a series of progressively less powerful axes which account for more nuance-like timbral characteristics as the axes account for smaller amounts of the variance. This leads to an higher-dimensional form than by attempting to explain all variance in a certain number of axes. Similar ideas can be applied to finding contributors to particular differences, such as that between muted and bowed strings.

The expected underlying characteristics of the timbre space can influence the way that research progresses. For example, if it is expected that timbral analogies ([123]) are a perceptually accurate frame of reference (e.g. A is to B as C is to D) then it might be expected that results from one area of timbre space may be translated to somewhere else. Or, as found in [123], such results may be accurate only in a limited context, or they may be direction-specific ([102]). At the other end of the scale, different sounds may be considered only a loose collection of points related in context-variable ways. In which case research effort might be placed on providing user control mechanisms to allow the human to adapt the sound appropriately through real-time feedback, rather than attempting to provide the user with fixed structures which can be relied on to provide the same effect in very different circumstances.

A particular problem which influences the results from timbre research is that it is impossible to apply a completely unbiased set of techniques. For example, given that it is known that attack rate influences timbre, to compare the attack rates of different sounds necessarily implies an algorithmic form which determines at which point the attack starts and ends, and which part of the attack thus defined is considered representative of the “true” value. It might be the average rate, or the fastest rate, or the average between 10 and 90%, and so on. When such decisions are applied to all specific measurements it becomes clear that two pieces of independent research into the same concepts are unlikely to arrive at exactly the same conclusions. Depending upon how sensitive the parameter is to algorithmic variation, such conclusions might be wildly different. Furthermore, it is hard to judge the robustness of a particular algorithmic form; it is infeasible to analyse the content of every parameter derived from every sound in a large set by hand for “quality”. There are then the problems of that which constitutes a representative set, by which algorithmic quality may be assessed.

Another aspect which influences viewpoint of investigation is the usage context. If the system must analyse, modify and synthesise timbre in real-time, then limitations are

placed on the time-dimension of investigation. For example, altering the length of the attack portion based on the amount of noise in the end part of the sound is impossible. Alternatively, in a non-real time situation, the feedback loop becomes long, which makes investigation of timbral change through experimentation harder. Another way of looking at context is the musical surroundings in which the sound form is investigated. If the timbre of note phrases is of interest, that is different to the timbre of single note instances ([146], [201]). Similarly, if the study concerns sounds within a group, then the interest is in what timbral aspects cause the sounds to blend in, or stand out.

In this research, the following usage aspects are important:

1. The aims are to investigate the structure of timbral perception through psychoacoustic testing, but also the underlying acoustical structure and how it distinguishes between different sound qualities and sound groups.
2. The acoustical structure is being investigated in terms of an hierarchy of axes, to attempt to understand nuances of sound as well as the larger aspects.
3. The parameters are based on previous researcher's forms, to an extent.
4. The system is not designed for real-time operation.

2.7 Dimensionality and Model Structure Concepts

The principles of perceptual organisation of timbre are intimately bound up with those of dimensionality and understanding both is necessary for appropriate investigation of its properties. There is often a desire to be able to deal with information in as few dimensions with as simple a structure as possible, to be able to display the data on visual media. However, the perceptual structure of timbre does not necessarily visualise well, especially when the number of dimensions involved is above 3.

2.7.1 Dimensionality

“All sensations exhibit a multiplicity of aspects, attributes, or dimensions, just as all stimuli are themselves multidimensional.” [198]

“The minimal number of axes required for an adequate representation of the data tells us how many dimensions are involved for a particular set of tones.”

[157]

Previous research has indicated that there is considerably more information (and thus dimensions of change) in acoustic signals than is necessary to describe timbral differences ([89], [66], [71], [137], [178], [201], [212]). This additional information may be of the order of 40-70% of the signal, according to that research. The component parts of the acoustic form are often significantly correlated, giving an amount of redundancy which allows the auditory cortex to find timbral cues from different parts of the sound ([80]). Thus, even if one cue is obliterated by noise, it may still be possible to achieve a reasonable estimate of the correct source from the other cues. However, some features may be highly localised, such as chuff at the start of sounds. Moreover, that as much as 70% of the data may be redundant still leaves a considerable amount of data that must be explained.

“Whilst there are results and ideas which indicate what acoustic aspects of different instruments contribute to the perception of their timbre differences, such differences are far too coarse to explain how experienced listeners are able to tell apart the timbre differences between, for example, violins made by different makers.” [88]

The interpretation of “redundancy” strongly influences whether the dimensionality of timbre is considerably less than that required to describe the acoustic signal. If the timbre space of interest ignores the nuances of playing style and concentrates on the major differences between instruments, then the redundancy within the signal is much higher than otherwise. To prove that a sample belongs to a particular broad region of timbre space is not a problem of very high order, compared to describing the more minor differences between samples from the same instrument. The highly complex information present in natural events facilitates enough features to be derived for the auditory cortex to establish a stable percept ([80]). This is particularly important if attempting to direct attention to a particular source among simultaneous sounds, for example.

A consistent mistake in the literature concerning timbre concerns the confusion between the dimensionality of a particular timbre space, and the number of dimensions necessary to

describe all perceptible timbres. In Chapter 6 of Plomp's "Aspects of Tone Sensation" concerning an experiment based on 9 stimuli he writes:

"In this example, based upon a specific set of stimuli, three factors alone appeared to be sufficient to describe the differences satisfactorily. This number cannot be generalized . . . It is also possible to select nine stimuli which would require, for example, five dimensions to represent their timbres appropriately."
[157]

This same text is interpreted by Hall as follows:

"There are, in fact, experimental indications that both vowel identification and musical timbre judgement depend on as few as three or four independent intensity parameters (Plomp, Chapter 6). This means it is at least theoretically possible that we may someday succeed in describing four specific aspects of timbre (or three or five - it is not all that certain) . . . with four such judgements together sufficing to accurately identify almost any timbre." [77]

Such a gross distortion of the available evidence is a common feature in the literature, and such concepts have gained acceptability almost by repetition. Stating that a particular timbre space can be described adequately by 3 dimensions is very different to stating that timbral perception in its entirety is a 3 dimensional construct. It is also exceedingly tempting to want to describe a complex multidimensional system in a dimensional (and structural) form which lends itself to graphical description, to make it easier to comprehend. However, there are other, more fundamental reasons why generally lower-dimensional timbre spaces have been the results of past experiments:

1. **The number of sounds and their range.** As described in previous sections, the number and type of stimuli principally limit the timbre space resulting from experiments. As described in Section 2.5, the number of sounds used has tended to be small. Moreover, the experiments have often investigated the large differences between sound sources and ignored more minor differences which would be present if the stimulus set included several examples from the same instruments.

2. **Other aspects of the model.** Similar arguments to those concerning the number and quality of sounds can be applied to all the other aspects of the model in use. In particular, the concept that only some parts of the sound are of relevance to timbre perception, such as harmonics and the steady state of the sound (see Section 2.9), has been used to limit the scope of consideration. For example, if it is only the steady state that is of importance, then all the considerable time variations which occur through sounds can be ignored. Reductions of this type also lead to confusion as to where variations come from. If noise bands cause masking effects in different sounds and the model is purely harmonic, then the perceived differences will be mathematically inexplicable. Whereas, if as much data as practicable is left in the representation, then there is the opportunity to find those differences in the data form. Similarly, the assumed structure of the model (see next subsection) implies how the nature of timbre perception is approached in terms of how nuances relate to major differences in quality. For example, whether a nuance of timbre is a major difference scaled down to a smaller perceptual level, or whether it corresponds to a different dimension entirely.

3. **The interpretation of “relevant” variance.** Experimenters often choose a solution space which represents the vast majority, rather than all, of the variance present in a stimulus set. For example, Plomp’s 3D solution accounts for 90.4% of the variance within the 9 samples. That leaves almost 10% unexplained in what is a very small set of sounds. Potentially a much larger set might be explained to the same degree, but as Plomp points out, a different set might well require more axes to represent it accurately ([157]). Similarly, Grey ([71], [72]) chose a 3D form due to its interpretability over a 2D or 4D form, not because it was a perfect fit for the original dataset. Preis maintains that one decisive parameter could be enough and all else are nuances ([168]). There are also implications for perceptual structure (see Subsection 2.7.2), in that the unaccounted-for variance may be considered noise; or alternatively important, but weaker, axes. Again, there is a question as to whether researchers are contemplating timbre as a description for all quality differences including nuances, or just the major differences between instruments.

4. **Simplification of the mechanisms of aural perception.** Although the sounds in use fundamentally limit the complexity of the timbre space, assumptions are also made about the way that the aural information is treated within the body. A

common example is the assumption that the critical bands of the ear are a means by which sets of frequencies are treated as pools of intensity, and thus only components which are in different critical bands are perceived as different ([77]). Thus the critical bands become a means by which the ear reduces the amount of data to about 15-18 simpler dimensions before the information is conveyed to the central auditory cortex. In fact, the purpose of the ear splitting the data into critical bands is to *preserve* as much information as possible, by allowing more data to reach the cortex through a number of neural pathways, when it can not all fit down one. It would be perverse were the auditory system to seek to reduce the amount of information reaching the central cortex, when the processing capability exists to consider aural information on a grand scale, such that minute nuances of timbre might be diagnosed. The more information available, the better the human's ability to understand and react to the complex World around it.

5. **The limitations of perceptual scope.** It is recognised that humans do not perceive all aspects of timbre at the same time, as the auditory cortex tends to concentrate on the most prominent aspects ([15]). As such, there are limitations to the scope of perception in a given context (see Subsection 2.7.2). There may be 3 or 4 dimensions perceived in one context, and 3 or 4 different ones in another. Or alternatively, the 3 most significant aspects may be constant, but there may be a group of less-significant axes which display differing balance of emphasis. Therefore it would be quite possible for an experiment based on a limited set of sounds to result in a form represented well by small number of dimensions; yet to represent all timbre in all situations with high accuracy might require a large number of axes.
6. **Desire for orthogonality.** It would be very convenient if the features by which timbre is judged in the auditory cortex had properties that were independent in perception and in the acoustic form. Non-interacting axes are generally easier to comprehend. Statistical techniques such as Factor Analysis (FA) and Principal Components Analysis (PCA, see Appendix B) exist in order to find orthogonal components which account for the majority of variance in a coupled system. However, having orthogonal axes does not necessarily imply that the system is easier to visualise than when it is described in terms of interacting axes. This is especially so with a system like aural perception, where the components of the acoustic form found to be of relevance (Section 2.9) have considerable couplings in their acoustic

and perceptual forms. It is easier to comprehend a system in terms of coupled axes which have a comprehensible form, compared to a mathematically decoupled form, comprising fewer dimensions, but having conceptually incomprehensible orthogonal axes. Sometimes the result can be a low number of dimensions with strong effects, and a group of smaller effects, which might otherwise receive individual treatment, on a residual axis. Grey's timbre space results can be interpreted in this way ([71]).

7. **Indirect evaluation through semantic scales.** The work of von Bismarck ([13]), Pratt/Doak ([167]) and others in determining an appropriate set of descriptors to apply to timbre is sometimes cited as evidence for the dimensionality of sound quality. Such techniques provide insight into how people describe sounds and which descriptions occur most often, potentially indicating the magnitude of effects within those stimuli. Unfortunately, language is always limited in its ability to express accurately, consistently and unambiguously such differences as those between a violin and a viola, or even two violins, playing the same note ([157]). The result is that such experiments are liable to result in a small number of common, well-defined descriptors and considerable confusion in the application of others, inevitably leading to a low dimensional conclusion.
8. **Interpretation by analogy.** Sometimes, the investigation of timbre has been subject to attempts to find parallels with other sensations. Many of the descriptions used of timbre are taken from those relating to other senses. In particular, visual perception is an interesting case. Padgham ([149]) compares the arrangement of organ stop qualities to a circular colour chart, where angle corresponds to "tone" and radial distance is "complexity". Such an arrangement implies an ability to perceive diametrically opposite timbre sensations and that all information can be conveyed in 2 dimensions. Pollard and Jansson ([165]) have drawn parallels between the action of the cochlea in the ear and the cone receptors in perceiving colour in vision, producing a form based on 3 features in 2 dimensions. The phenomenon of synaesthesia is occasionally mentioned ([50]) and cross-modality matching has been demonstrated ([198]), which could lead to an ability to systematically compare values in one sense with another to some degree. However, while interesting comparisons can be made with other experiences and particularly other sensory perception such as vision (see Subsection 3.3.1), it can be misleading to make comparisons that are too direct. Almost inevitably with a complex and badly understood phenomenon

such as timbre, the result is unlikely to be a correct assessment of the dimensionality involved. With [149] and [165] it would seem that the result is two dimensions, but the comparison is based in only a small part of the acoustic form (they both limit consideration to harmonic information). Yet it would be equally feasible to compare timbral entities to, say, the properties of physical objects as perceived in the mind; which includes concepts such as structure, apparent texture and weight, classifications of usage, familiarity and so on. That could result in a very large number of axes; there is no fundamental reason why human vision should comprise very complex interacting dimensions, and that timbre is perceived in a low number of dimensions. Yet Padgham, Pollard and Jansson chose low dimensional constructs, based on preconceived ideas of dimensionality.

Despite the above points, timbre may yet be reduced to 3 or 4 fundamental dimensions. However, there has not been enough evidence to show that it is the case. Crucial to this argument is the number and qualities of sounds used in experimentation, and stating exactly what timbre space is actually being considered, such that different experiments can be compared. For example, it would be poor practice to consider timbre the “psychoacoustician’s multidimensional waste-basket category” ([122]) and then proceed to ignore a considerable proportion of the variance between sound qualities, when that proportion accounts for the nuances, which are necessarily part of such a loose definition. Chapter 5 provides evidence to promote the concept that the auditory cortex uses a considerable range of acoustic features and that perception may concern a considerable dimensionality extracted from those features. Another way of looking at the dimensional problem is to consider whether it is likely that timbre is actually a relatively simple phenomenon, merely being dealt with in a complex manner in the human hearing system (as indicated by the physiological anatomy of the ear and psychoacoustics). This seems unlikely. Moreover, the human cortex is fully capable of dealing with exceedingly complex operations, such as manipulating tools and playing instruments based on sensory feedback. There is no implicit reason why timbre should be a simple sensory construct.

Interestingly, in [212] it is suggested that increasing the amount of variance in assessed timbres does not necessarily lead to proliferation of the number of dimensions required to represent the timbral differences. This is based on the assumption that high level descriptions are common to all sounds. It is hard to prove conclusively that more and

more dimensions are required to cope with an increase in the different timbral nuances used as input to a system, or whether there are common timbral descriptors for all sound types, without considering a very large number of input stimuli. It seems likely that there is a limited number of axes of perceived timbral difference, where that number is not small enough to be easily visualised or incomprehensibly large. However, what is known is that if dimensionality is reduced too far, there is some evidence that the metrical relationships between the input stimuli is compromised ([200]).

It is reasonable to consider that complex phenomena can be represented in low numbers of dimensions. Statistical techniques such as multidimensional scaling (MDS, see Appendix B) allow considerable dimensionality to be optimally reduced to a low numbers of dimensions, at the expense of absolute accuracy. The auditory cortex might deal with acoustic information in a similar way. Krumhansl ([102]) is sceptical. Given the vast parallel processing capabilities of the brain, it seems more likely that a scheme by which a large number of individual features are tested for fit and the final perceived result is a combination of the answers from those different parts would be used. A large scale reduction to a simpler form seems less likely. This raises the question of whether the aim is to match the engineering model under construction to the cortical model; or whether the aim is actually to construct an effective description of the timbre perception process, where the result has as simple a form as possible. The answer is probably the latter; but that doesn't immediately imply that the the engineering model will have only a couple of dimensions, merely that it is likely to have fewer dimensions than the cortical equivalent.

A dimension of timbre might represent a very complex spectral transformation process and thus encompass more than one simpler dimension. However, the question is whether investigations aim to create the most mathematically compact timbral model forms possible, or whether developing a form with a number of *comprehensible* feature dimensions is the priority. The latter seems likely, although it depends upon the application of interest; the requirements for data visualisation may be very different from those required in musical performance control. Experiments concerning dimensionality are detailed in Subsection 5.6.3.

2.7.2 Model Structure

There are a considerable number of different areas of the spectral form which researchers have found to have importance (see Section 2.9), which must somehow fit into the model in use. One method of dealing with the structure of timbre space is as a plain multidimensional form where the position of all sounds is determined on all axes at once. However, it is possible to structure timbre space such that specific variation is accounted for within an instrument's subspace in a more progressive manner. The necessity for such structures is indicated by the limitations of low dimensional solutions. Such solutions account for a considerable proportion of the variance between sounds (as explained in the previous subsection), but fail to explain "unique factors"/nuances associated with the sounds, indicating further detail at a lower level. Such aspects are sometimes referred to as "specificities", explaining such things as the bump at the end of harpsichord tones, or "hollowness" in timbre ([43], [102], [123]).

One way of dealing with specificities is to consider them add-on components which represent that which makes the particular sound recognisable. Thus, every sound can be located in a general way in a low-dimensional timbre space, and then needs its uniqueness factor to distinguish a particular sample. Alternatively, there might be another layer; a uniqueness factor to locate the instrument type, and then a set of factors to locate particular samples within the timbral subspace of that instrument. This can be viewed as context-dependency, which is implemented in some systems ([212]). A particular advantage of the specificities plus basic axes timbral structure form is that it easily explains complex characterising aspects like the onset "blip"s found in trumpet tones ([138]) without having to take a systematic and generalised approach to characterising these forms. For example, to characterise the attacks of all sound sources in terms of coherent axes rather than a collection of specificities, it is necessary to be able to extract features that explain all the different types of attack.

Specificities represent one solution to inaccuracies in the low-dimensional forms, but an alternative is to consider timbre as an hierarchy of embedded distinctions ([102], [111]). That is, that timbre space can be navigated through a series of axis sets with decreasing strength of perceptual difference. For example, the region of timbre space occupied by woodwind and brass can be found from a major axis set; then a more minor set used to

approach the trumpet-like region; then the more minor set to find a cornet; and finally the most minor set abstracted from the previous levels to navigate the instrument space. This is shown in Figure 2.4, where each oval represents a region of timbre space, within which movement is described by an axis set, dependent upon having reached that region by the feature set describing the region above it.

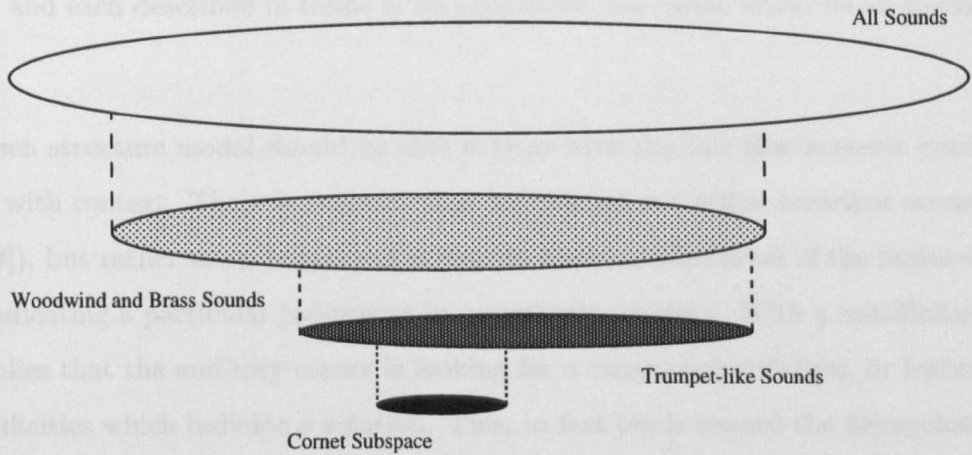


Figure 2.4: Example Hierarchical Structure of Timbre Subspaces Described by Axis Sets

An hierarchical representation is not only fine for describing the processes of focusing in upon a particular instrument type, but is equally valid for the description of general timbral regions which cross instrument set boundaries (Figure 2.5).

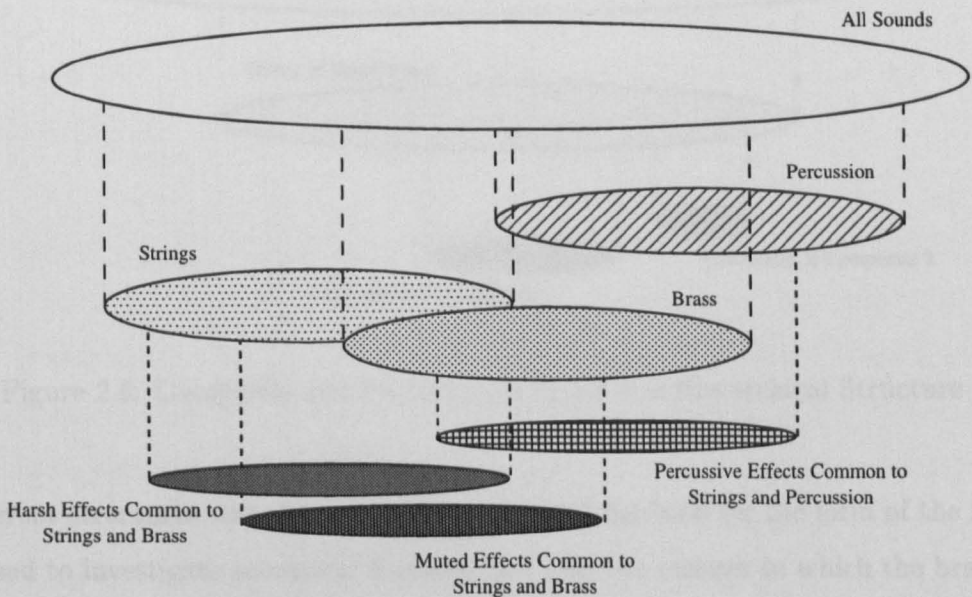


Figure 2.5: Example Hierarchical Relationships of General Timbre Types

These concepts tie in particularly well with the processes of sound production from physical sources, which have an additive form of components from different parts of the resonating source, which have different levels of contribution to the overall sound; an hierarchical combination of sound properties ([80]). Acoustic signals in general can be considered an infinite series of correlation functions ([46]). If those were ranked in order of salience, and each described in terms of its properties, the result would be an hierarchical form.

The chosen structure model should be able to cope with the fact that acoustic cues are variable with context. There is evidence that humans do not utilise invariant acoustic cues ([80], [89]), but rather use a judgement structure based on likelihood of the features of the sound indicating a particular judgement in a particular context. With a specificities form, that implies that the auditory cortex is looking for a range of specificities, or features of the specificities which indicate a solution. This, in fact tends toward the hierarchical view whereby variations within a classification region are described by the lower level axis sets. Handel ([80]) suggests that representations may be composite when sound instances are similar, but there may be different prototypes when the differences are great. An example of this is given in Figure 2.6

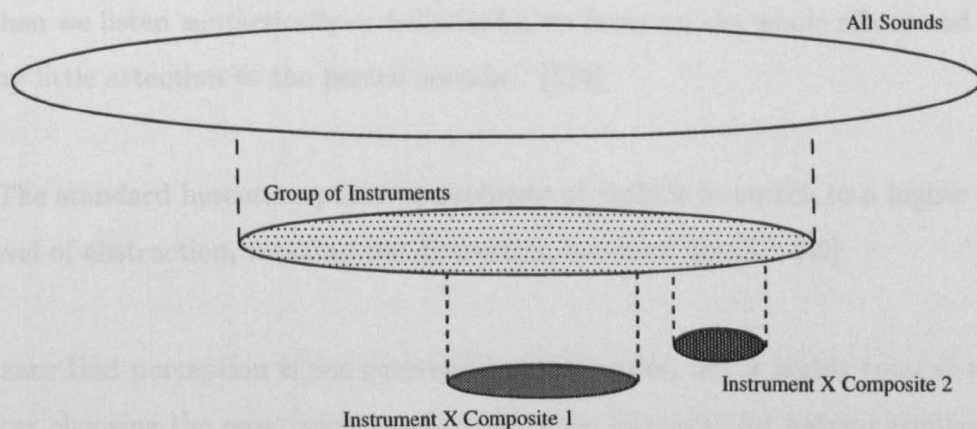


Figure 2.6: Composite and Prototypical Splits in a Hierarchical Structure

The different structures and viewpoints have implications both for the form of the research model used to investigate acoustical features, but also the manner in which the brain is assumed to structure timbral information. Given the current state of knowledge, no single form is known to be correct. Furthermore, the most appropriate representation cannot be determined from a few samples. In the near future the computing power will be widely

available to test such theories by examining a very large number (many thousands) of acoustic forms containing many examples of each of the sound sources of interest. Only then will it be reasonable to draw sweeping conclusions about how sounds are best represented in timbre space and what the axes/specificity types are that describe the differences.

It is within an hierarchical framework that this research develops its theories of timbre. It is used, in particular, as it is a systematic approach to characterisation which allows analysis in a progressive manner. It relates well to knowledge of perception:

“There is a psychoacoustics speculation that humans tend to organise their knowledge in layers.” [133]

Patterns of information are considered on several levels in the human brain ([5]). Secondly, perception of timbre is a scalable phenomenon:

“Our auditory system has the ability to listen to complex sounds in different modes. When we listen *analytically* we hear the different partials separately; when we listen *synthetically* or holistically, we focus on the whole sound and pay little attention to the partial sounds.” [174]

“The standard human response to problems of scale is to switch to a higher level of abstraction, masking the distracting low-level detail.” [40]

This means that perception is not governed by static rules, but is highly context-sensitive; the cortex choosing the most appropriate level in the hierarchy for judging similarity ([15]). This concept is reinforced by studies that show that if listeners are asked to judge timbral differences between a wide range of sound qualities, then they are likely to focus on the most significant perceptual axes of difference and ignore smaller changes. Whereas, if the task is to consider the differences within a group of similar sounds, subtle axes of difference become important, which were swamped when considering a larger timbre space region. For example, in the perceptual study in this research (Chapter 3) a number of observers noted that they thought some of the stimuli being judged were repeated within each test, when in fact the sounds were all different. This occurred because some of the

sounds had small degrees of difference (such as bowed violin and viola notes) which are very recognisable when are compared in isolation, but not when considered in the context of the whole group.

2.8 Overview of Research Techniques

“The most distinguished physicists, when they attempt logical analysis, are apt to gibber; and more nonsense is talked about measurement than any other part of physics.” [23]

There are different ways of approaching research into timbre. The aim of such research is to identify dimensions which are equivalent to those used in the auditory cortex to distinguish between sound qualities. Or, to look at it in a slightly different way:

“Research on musical timbre typically seeks representations of the perceptual structure inherent in a set of sounds that have implications for expressive control over the sounds in compositions and performance.” [218]

The search for the dimensions of interest may occur from either a perceptual and/or an acoustical viewpoint; timbre perception combines both perceptual processes with acoustical properties ([80]). In general (and in this research) the purpose is to relate the two together ([46], [197]), although some research is only concerned with the psychological rather than the psychoacoustic ([130]). Structures which explain timbral form can be developed from both perceptual and acoustical experiments. As explained in Section 2.7, the aim is to generate a structure from experimental data which conveys an interpretable form, such that meaningful conclusions can be drawn. The ability to achieve this depends upon the techniques used to analyse the sound properties in the first place. Three major classes of techniques have been principally used in timbre, as follows.

2.8.1 Evaluation of Timbre by Semantic Scales

Studies which have considered how to describe sounds along semantic scales (based on perceived physical attributes such as “hollow”, “heavy”, “bright”) have produced a deeper

understanding of the ways in which humans describe sound. Some authors consider that descriptions provide handles on timbre that are highly comprehensible ([194]). However, a difficulty remains in identifying acoustical correlates with descriptive scales. Knowing that the differential scale “bright-dull” is an important opposition when describing sounds does not indicate an acoustical source, unless the parameter that controls perceived brightness can be found within the acoustic form that correlates with the described changes. Indeed, concentrating on that which can be described can be a dead end, in that many timbral concepts are not easily described by semantic scales, or words may describe several effects, be interpreted in different ways, or may be medium-specific ([13], [56], [80], [93], [149], [213]). This is not helped by the fact that timbre encompasses a very diverse range of effects.

Different terms do not display the same degree of consistency in interpretation; Waters ([213]) notes that those which deal with processes (“looping”, “accumulating”, “accelerating”) are less problematic than the descriptions which pertain to attributes of physical objects (“wide”, “hard”, “rough”). Another problem is that asking subjects to rate sounds along those scales imposes preconceived ideas about the description of sound. For example, that the descriptions are specific enough to prevent confusion, and that a group of descriptors can be assembled which have sufficient coverage to describe all presented timbral relationships ([171]). The lack of a unique relationship between semantic scales and independent properties of stimuli is pointed out by Grey ([71]). There is also the suspicion that when rating sounds on a multitude of axes the most salient quantities swamp perception and make correct evaluation of lesser aspects much harder.

A distinct problem is that an inability to describe an effect explicitly in words does not imply that an effect is lacking in perceptual solidity, or that in usage as a variable modification to sound is incapable of being intuitive. However, considering in perceptual terms avoids the problems associated with trying to evaluate the appropriate properties of the acoustic form directly:

“... a musician’s description of timbre is not equivalent to a description of a sound’s frequency spectrum” [228]

A considerable number of authors list “typical” descriptions, for example:

“*soft, piercing, braying, hollow or poor, full or rich, dull, bright, crisp, pungent* and so on.” [83]

“Texture always denotes some overall quality, the feel of surfaces, the weave of fabrics, the look of things. Words from visual and tactile sense modalities are often appropriated for descriptions of sounds and their combination: sharp, rough, dull, smooth, biting, bright, brilliant, brittle, coarse, thick, thin, dry, diaphanous, airy, finespun, flaccid, fluid, gauzy, glittery, grainy, harsh, hazy, heavy, icy, inchoate, jagged, limpid, liquescent, lush, mild, murky, pliant, relaxed, rippling, to name a few.” [51]

Other authors conducted more empirical work. The following are of particular note:

1. von Bismarck ([13]) considered how 30 different pairs of antonyms in 4 groups (based on the relationships to the most important factors hard-soft, compact-scattered, empty-full and colourless-colourful) could be used to describe a group of sounds. Plomp ([157]) says of this that semantic scales reveal strong factors, yet are limited by their inability to describe all relevant nuances of sound (see also [171]).
2. Solomon ([196]) considered 50 descriptive oppositions which were used as rating scales for sonar noises. These appeared to group together into magnitude-related, aesthetic-evaluative, clarity, security, relaxation, familiarity and mood.
3. Pratt and Doak ([167]) attempted to find important descriptions of timbre and whether they could be consistently applied and used to provide quantitative estimates of factors governing timbre.
4. Ethington and Punch ([52]) have achieved more than most in creating a system to actually test some of their semantics, by relating acoustical parameters to the descriptions. These were identified in the three classes of attack, presence (sustain) and cutoff (decay).
5. Nordenstreng ([139]) notes that the descriptive factors applied to speech sounds and “musical” sounds are not necessarily the same.

Other work considering timbral description includes [12], [99], [105], [130], [142], [197] and [215].

2.8.2 Evaluation of Timbre by Perceptual Distance

“... the ability to generate a data structure on the sole basis of perceptual judgements, which may then be interpreted with respect to the physical parameters of the sounds, is especially appealing with stimuli so complex as natural timbres.” [71]

An alternative to perceptual rating of sounds along descriptive scales is to rate sounds for perceptual disparity or similarity. This is a common technique for quantifying the relationship between sounds, as it is not concerned with semantics and the results create a timbre space which is not predisposed to a particular analysis model.

“With few exceptions, we tend to judge an object in terms of the relationships it has with other objects.” [218]

A major assumption with such work is that those asked to judge similarity between sounds are utilising all the attributes that govern sonic perception ([198]) and are not concentrating on particular aspects. This is difficult to guarantee, both with sonically-aware¹ and generally untrained subjects. Also, there is a difficulty in judging the (dis)similarity of very different sounds accurately (such as the distances between a piano, a car engine and bird song) compared to the accuracy possible in judging similarity between, say, string instruments ([13]). A particular problem with analysis of perceptual judgements is that they are prone to chance error ([80]), which may cause a shift in the resulting dataspace, which is very hard to identify. Furthermore, the problem of matching dimensions derived from a perceptual space to acoustical axes of models of sound still exists ([62]). Thus judging similarity gains great insight into the structure of timbre perception, rather than understanding directly how perception is linked to the acoustics of sound.

The following studies are of particular note:

1. The best known study of timbre spaces based on perceptual distances between sound sources is Grey's thesis ([71]). Grey asked subjects to compare sounds equalised in pitch, loudness, and duration for perceptual similarity. Those distances formed a

¹Those persons who are experienced in the techniques of listening and analysing sound

multidimensional space that was reduced to a small number of dimensions by multidimensional scaling (MDS, see Appendix B) . Three dimensions was found to produce an interpretable solution relating to spectral energy distribution; low amplitude, high frequency energy in the attack (possibly inharmonic), potentially implying “hardness” or “explosiveness” of attack ([74]); and synchronicity of attacks and decays of higher harmonics (or possibly musical instrument family relationships).

2. Wessel ([218]) considers the data in a similar manner to Grey, producing axes of brightness (to do with spectral energy distribution) and bite (onset characteristics). Iverson and Krumhansl ([93]) had similar results of brightness and percussiveness.
3. Plomp ([157]) points out that, in effect, judging perceptual distances is similar to describing sounds in terms of others - their relationships within a reference framework of known sounds.
4. Saldanha and Corso ([175]) considered instrument identifiability under different conditions.
5. In Donnadieu et al’s perceptual test ([43]), similar axes to other authors of “brightness” and “attack quality” were found, but the third dimension seemed to relate to specificities (see Section 2.7).

Other work concerning perceptual distance includes [19], [62], [93], [105], [120], [131], [142], [149], [154], [201], [217] and [218].

2.8.3 Evaluation of Timbre by Acoustical Parameters

Starting from an acoustical or physical description of sound (such as a spectral, or physical model) has the potential to be as limiting as starting with perception, because the ways in which features are extracted from the model form implies preconceived structural forms ([71]). However, there is the possibility of creating a general description of sound, from which a large supply of descriptives might be abstracted:

“Mathematically, an acoustic signal can be measured in an infinite number of ways. Psychological research in timbre helps to define which mathematical properties of sound are relevant to human timbre perception.” [107]

Inevitably, it is possible to take the opposite view; that the open-ended nature of a system which allows an unending supply of descriptions is one which is prone to never fully characterising the system. The counter argument is that, through perceptual examination, it is possible to establish those areas of the spectral form which are of interest, and then concentrate on finding the particular parts of those areas which are of importance by statistical evaluation and hierarchical decomposition. Starting with a plausible acoustical model and then searching for perceptual meaning is sometimes called a reverse research methodology ([46]).

By considering different perspectives on the spectral form which are not necessarily orthogonal, it is possible to establish the important areas. Previous research has established those parts which are consistently found to be of perceptual relevance. However, a particular problem is dealing with perceptual interactions between parameters, and thus finding appropriate *sets* of features to describe a timbral difference, rather than expecting a simple single parameter to be enough. Also, there is the possibility of considerably different model conditions leading to similar perceptual states ([15]), and small differences in models leading to very different perception. However, the major benefit of considering acoustical parameters over purely perceptual experiments is that it is possible to directly measure and manipulate data.

The following studies are of particular note:

1. Lichte ([112]) considered variations to spectral balance and harmonics, which were then related to descriptions.
2. Webster et al ([214]) considered how different patterns of harmonics relate to identifiability and thus timbral differences.
3. Pollard and Jansson ([165]) investigated the development of groups of harmonics (fundamental, harmonics 2-4, harmonics 5-n) to better understand the form of the starting transient behaviour and its relationship to the steady state portion.
4. Slawson ([189], [190], [191]) considered the effects on timbre produced by varying formant positions for vocal-like sounds, generating the dimensions of openness, acuteness, smallness and laxness.

5. Ethington and Punch ([52]) took variations in acoustical parameters and attempted to relate the perceptual effects produced to those listed in their set of 124 timbral descriptives. Problems such as confusion between variations and complex choices between perceptual categories existed.
6. Wishart ([226]) considered a considerable number of techniques for varying sound qualities, which have known effects, and whose usage can become intuitive.
7. Langmead ([107]) considered a group of 9 parameters extracted from a spectral model known to be related to timbre, to investigate how well their dissimilarities describe timbral differences. This was subsequently increased to 14 parameters ([106]).

Rather than considering individual acoustical components, some authors have taken a larger description and reduced the number of parameters to a representative set of lower dimensionality, using procedures such as principal components analysis (PCA), MDS and other scaling techniques (see Appendix B for details). These attempt to find the fundamental components which represent the differences between the input stimuli. Such statistical methods have also been used with the other research techniques.

A different class of techniques for reducing a considerable set of acoustical parameters to a more compact representation, is to use self-organising maps (neural network techniques). This generally involves extraction of salient aspects of sounds from an acoustical representation, which are then used to train the map to find the most important dimensions ([34], [56], [160], [201]). It would be possible to train the system to identify particular timbral aspects such as “mutedness” or “woodwind-like”. There is considerable evidence that the topology of the map can be related to the topology of a timbre space derived from similarity judgements ([201]). The problem with such a technique is that, while it mimics the learning style of the human cortex to an extent ([201]), it fails to aid the researcher in knowing which particular parts of the original representation are important in as simple a manner as the weights derived in a PCA system. Other authors who have considered the use of neural networks in timbre are [107], [133], [132], [152], and [228].

2.8.4 Technique in this Research

There are often considerable variations in the approaches of different investigators, and it is also apparent that no single technique or type of technique has a monopoly on the path to truth about timbre, or is most simple to implement. This research considers a two-pronged attack, to achieve both direct measurement of acoustical features which correspond to important timbral areas, and also perceptual testing to better understand how such features relate to the aural experience. This combination consists of a perceptual study involving the judgement of the timbral similarity of stimuli to regions of timbre space (Chapter 3); and measurement and analysis of features extracted from a timbral form for the stimuli of interest (Chapter 5).

Direct measurement of spectral features enables extensive statistical analysis and interpretation of known salient characteristics. If desired, it is also possible to take subjective assessment out of the process by using features derived automatically from the acoustical description ([162]). However, as timbre research is concerned with perception, it is unreasonable to attempt to totally detach assessment of timbre from its perceptual origins. From the other perspective, judged similarity data enables interpretation of the principal axes of the timbre space, and comparison with acoustical attributes which correlate with them. Authors have often combined techniques and viewpoints in similar ways to this, to aid in understanding.

This approach avoids consideration by means of verbal rating scales. The role of semantic scales in understanding the perceptual nature of timbre is taken by the similarity study. The problems associated with such rating scales are often considered overly difficult in comparison:

“... the use of verbal rating scales to classify timbres is a very precarious if not a dubious tactic.” [71]

2.9 Important Aspects of the Spectral Form

Over time, a considerable number of authors have attempted to establish the aspects² of the time-varying spectral model relevant to timbre. Although conclusions from these experiments are sometimes conflicting, the fact that relationships between acoustical features and perception are consistently found, indicates that these are, at least, important parts of sound. They are certainly not an orthogonal set nor equally important in all circumstances.

1. **Onset / prefix / attack characteristics.** The onset of a sound indicates the method of its inception ([226]) and the auditory cortex pays great attention to pattern onset in sensory information ([5], [150]). Its importance is indicated by the fact that without its presence, the ear can find it hard to distinguish sounds which otherwise might be considered clearly different ([71], [172]). Furthermore, altering the attack can radically shift the perceived source characteristics ([226]). Considerable variation in attack can be achieved through playing style, as well as that found between instruments ([51]). It is noted that onset refers to both the amplitude envelope and the underlying spectral information ([71]). With traditional instruments, the interference of free vibration, forced motion and background effects form the starting transients ([164]). Other references which consider onset include [11], [15], [24], [22], [43], [44], [53], [77], [88], [91], [93], [100], [102], [106], [109], [118], [123], [126], [127], [136], [137], [156], [165], [171], [175], [177], [199], [201], [200], [212] and [219].
2. **Steady state.** The wealth of studies considering what “steady” characteristics may be found testify to the importance of “general trends” within sounds, sometimes considered the “intrinsic” properties ([77]). Early authors ([83]) considered steady aspects the most important parts, but more recent work tends to put a similar emphasis on both onset and “steady” parts ([171], [173], [176] and others); steady state alone being insufficient for clear instrument identification between all sounds ([77]). The concept of steady state is somewhat vague, though, as only a small proportion of sounds have steady characteristics within the length of the sound,

²These general areas of the spectral form of interest will be known as “aspects” and particular parameters derived within the areas will be known as “features”

except on a very small scale or in terms of overall trends ([88]). As such, the existence of a “steady state” has been questioned by some authors ([51]). However, because timbral properties are varying over time, it is often necessary to consider an average effect, or trends within the data, to make the large quantity of information accessible. Pollard and Jansson ([165], [163]) show how the steady state is a region of timbral variation, with different notes on the same instrument also characterising an area of timbre space. What is important is that there is generally found to be a significant divide between the initial part of the sound and the remainder. The remainder is generally the region with more steady characteristics. Other references which consider the steady-state include [15], [19], [52], [149], [168] and [206].

- 3. Decay.** Like the steady state, it is difficult to consistently delineate a particular part of all sounds which might be considered to be “the decay”. Often, the whole of the sound from the end of the onset is a decay. Again this is a convenient description on a broad level, but difficult to characterise. The concept of decay has additional confusion in that it may refer to the general decline of amplitude (usually from the end of the attack portion), or after the excitation is removed (the sustain or “steady” state) from the sound-producing source (assuming a physical source), which is also referred to as the release or offset. The fact that the timbre of a sustaining sound can be perceived before the final offset indicates that the offset is not as important as the onset in determining timbre in that situation ([88]). However, if the sound is short and non-sustaining then the offset has greater importance. Other references which consider decay include [52], [109], [164], [173], [177] and [199].
- 4. Amplitude envelope.** Treating the whole envelope as an object with overall defining properties is sometimes a more profitable way of examining sounds than by looking to split the sound into sections. Sounds from physical sources are always more complex than just an attack, sustain and decay ([136]). The overall envelope indicates much about how the sound source is forced to develop (the articulation involved, [193], [211]). A struck sound, for example, will have a fast attack, but a instrument which is struck and then resonates is different to one without continuation. Similarly, a blown instrument has a continuous excitation, whereas a struck one does not (forced and dispersive continuation respectively). Attention can be drawn to different parts of the sound based on the morphology ([193]). The development of amplitude can be strongly linked to the underlying spectral

information, and has an equivalent importance ([226]). Other references which consider the amplitude envelope include [15], [51], [93], [107], [106], [200] and [212].

5. **Harmonic form.** The human aural system has a tendency to search for harmonic relationships. They indicate important aspects of a resonating source. Such forms then, are particularly prevalent in the acoustic form of traditional musical instruments rather than “everyday” sounds ([67]). This also leads to extensive coverage in the literature. There is a danger in placing too much emphasis on harmonic relationships compared to other components of the spectral form, however ([77]). Different authors have assigned different meanings to individual harmonics (particularly up to about the 6th or 7th), or groups of harmonics. The degree of match to the strict harmonic template of multiples of a fundamental frequency (sometimes known as “harmonicity”) has also been shown to alter timbre. The distribution of emphasis in different harmonics or sets of harmonics is often considered important. Other references which consider harmonics include [7], [15], [22], [46], [52], [88], [83], [96], [100], [107], [109], [112], [127], [131], [137], [147], [149], [161], [165], [167], [168], [173], [175], [176], [180], [201], [206], [211], [212], [214], [215], [217] and [226].
6. **Inharmonic form.** In the presence of harmonics, a range of effects have been noted as regards inharmonicity, from a light effect increasing “warmth” or more “natural” characteristics, to larger effects producing roughness at higher frequencies ([51], [118]). However, with a source which displays little harmonic information, it is the inharmonic structure itself, rather than a relative effect that is important. “Inharmonic” does not imply a specific form, and the range of effects can depend on such things as inharmonics’ amplitudes and clustering relative to neighbouring spectral elements. Often, inharmonics have been considered a category which is simply “everything else” when a timbral model is based on harmonic effects, when in fact it encompasses a wide range of spectral forms. In the modern context it is unreasonable to lump everything that is not harmonic of the fundamental frequency into a catch-all bracket. Other references which consider inharmonic form include [7], [63], [71], [112], [98], [127], [131], [147], [164], [171], [172] and [226].
7. **Patterning of frequency information.** The above two items may be considered to be part of a general distribution of data within the spectral form. Lundén ([116])

splits the data into a discrete section, containing the relationships between harmonics and partials in general, and a continuous aspect, which implies bands of information (“noise”). Many other aspects of patterning could be considered in a more generalised way than they have to date (having mainly been concerned with harmonic patterning). Proximity effects are often mentioned. Other references include [7], [15], [30], [88], [157], [171], [174] and [214].

8. **Spectral contour.** The large-scale characteristics of the spectral form are consistently mentioned as important as they indicate the resonant structure and emphasis/spread of information in the spectrum ([30], [51], [71], [173], [188], [190], [218], [226]). However, the form which this takes depends upon the instrument; sometimes an approximately fixed resonant (formant) structure (particularly with vocal sounds), sometimes a pattern of amplitude which changes in particular ways with loudness, pitch and so on. Sometimes consistent resonant patterns or trends are not present. The overall spectral balance is often mentioned as an important feature, often through the spectral centroid, or frequency spread, or the spectral slope. Contours may be static with time, or show considerable movement, which imply particular source characteristics. It has been suggested that the principal parameter determining stimulus timbre could be derived from the shape of the spectral envelope ([168]). Other references which consider the spectral contour include [1], [6], [7], [15], [19], [24], [22], [43], [44], [77], [80], [98], [100], [102], [105], [107], [106], [115], [123], [137], [138], [150], [157], [172], [175], [176], [184], [185], [189], [191], [210], [211], [212], [217], [219] and [225].
9. **Noise aspects.** Noise is loosely broad bands of frequency information, or irregular vibration, with considerable energy in this context. It is important to note that noise is not simply “white” over a large bandwidth, but can have different forms of filtered dense or loose clustered partials ([226]). This results from there being a relationship between the elements of the spectrum in natural processes, rather than the bands of partials being governed by purely random effects ([25], [128]). Some authors consider there to be a noise-note continuum of sound types ([51], [179]), which may be linked to ability to resolve pitch information ([193]). Noise can result from very different effects, such as the passage of air through a wind instrument, natural irregularities in the system, mechanical noise, interfering resonances, or recording effects/environment ([64], [83]). Sometimes such aspects are considered essential to the

timbral nature of the sound, but some authors regard them as environmental or recording anomalies, depending upon the type of effect. Particularly prevalent is noise at the beginning of sounds, where inertia of physical systems often produces considerable frequency spread ([51]). Also, the type of excitation that the system receives (such as blown, or bowed, or struck) results in different noise patterns. Indeed there is no clear line of demarcation between what is “tonal” and that which might be considered “noise”-based ([7], [50]), although some analysis-synthesis systems decompose the spectral form into stochastic and deterministic parts to enable the information to be dealt with separately (for example, [124], [181]). Other references which consider noise aspects include [15], [30], [80], [94], [116], [164] and [172].

10. **Temporal evolution.** Perception concerns temporal as well as spatial dimensions ([5]). This includes features from the large scale movement of overall spectral contour, to harmonics and inharmonics in frequency, amplitude and (sometimes) phase. The change (instability) in the spectral form through the sound is sometimes considered a fundamental aspect of timbre, described as spectral “flux” or “fine structure” or “irregularity” ([43], [100], [102], [123]). Also, aspects which are in motion can themselves have a time-evolving characteristic which is perceptible ([226]). Chowning ([29]) considers the character of temporal evolution of spectral components “critical” in determining timbre. If the model representations of sounds are to be manipulated with respect to each other, then having handles on the significant parts of the evolution is vital in controlling the sound without losing its nature ([115]). It is not obvious, however, whether it is always the presence of evolution, or the form (shape/speed/order) of evolution, or both that are important ([93]). Sometimes, the temporal aspects may be considered a separate domain in their own right ([116]). It is apparent that a number of timbral effects derive from temporal proximity and persistence, and the associated fusion and strength of percept ([226]). Temporal evolution may be considered at very different levels, as well, such as micro-intonation in the sound, or a large effect, like a vibrato. Other references which consider temporal evolution include [1], [15], [46], [25], [30], [47], [51], [63], [68], [71], [77], [80], [107], [106], [117], [127], [136], [137], [138], [150], [161], [164], [171], [172], [173], [174], [179], [180], [201], [200], [211], [212], [218], [222] and [224].

11. **Transient (momentary, short duration) aspects.** Transient aspects particularly include features which typify “natural” sounds and are hard to predict, and thus are hard to analyse in a generalised way. This includes mechanical clicks, momentary noise from performance mechanics (transients in bow action, fingers slipping over holes), inequalities in air flow and so forth. These are particularly likely to occur at the beginning and end of single sound instances. References which consider transient aspects include [25], [51], [64], [77], [83], [138], [164], [172], [174], [175] and [217].
12. **Synchrony of partial movements.** Some authors note that the movements of partials within the spectral form as regards their (a)synchrony in different parts of the spectrum, or with relation to other parts of a pattern of partials (such as the harmonics) can have an important effect on the way the sound is perceived to be produced. Temporal order can be important in recognition of time-varying patterns ([5]). Onset and offset asynchrony is often mentioned, although onset asynchrony is the more detectable ([47]). Other references which consider synchrony include [22], [25], [51], [71], [80], [107], [138], [161], [165], [173], [175], [211] and [218].

The main divide apparent in the description of important spectral components is along the lines of frequency. Those aspects which have very low frequency are often described as being “temporal evolution”. This includes the overall tendencies of amplitude, frequency tracks, vibrato and so forth. Such components are often considered separate from audible rate phenomena. There is evidence to suggest that there is a perceptual divide present ([67]). Furthermore;

“The majority of neurons in the auditory cortex are unable to signal envelope modulation at modulation rates of much more than 20Hz” [150]

2.10 Specific Features of the Spectral Form

Section 2.9 discussed the general aspects of the spectral form which have been found to be timbrally salient. However, some authors have been more specific in their assessment of the spectral form, giving specific features of the salient aspects. These were not necessarily chosen because they are the most salient out of a range of possibilities (although in some

cases that is true). In different timbre spaces, they were observed to display a correlation with timbral effects. There are other features which have not figured in experiments to date. But, the list covers a wide range of spectral forms, which provide a basis for further examination.

Whereas the spectral aspects in the previous section have substantial experimental backing to testify to their general importance, specific features such as those listed here are necessarily more specific to the timbre space within which they were observed. What is apparent from the list, however, is the sorts of methods and feature extraction forms which have been studied and found successful. These indicate the sorts of places to look for timbral information. It is also noticeable that there are substantially fewer specific forms than references to general areas in the previous section. This is partly due to duplication and repetition of results, but also that many results have been qualitative or non-specific. There is no particular way of measuring “attack quality”, or “random aspects”, or “inequalities of acoustic form” for example.

The list contains features which are more associated with general application than with particular specific sound sources. Also, this is a list of features not methods of processing data. There is a very large number of methods of altering the spectrum through different techniques (see [26] for example), but here the interest lies in what components of the spectral form are directly important. However, due to the generally unspecific nature of timbre studies to date, some of the features listed below are also somewhat ambiguous in their form. Features which relate to specific semantic descriptions are given in Section 2.11.

1. Onset / prefix / attack characteristics.

- (a) Rate/Length of onset. [15], [24], [51], [80], [83], [93], [100], [102], [171], [172], [173], [123], [127], [131], [201], [226]
- (b) Amount of energy at the start compared to the rest of the sound. [226]
- (c) Presence/quantity of noise/inharmonics. [123], [127], [171]
- (d) Frequency spread in the attack. [51], [127]
- (e) Low amplitude, high frequency (inharmonic?) energy. [71]
- (f) Beginning, end and peak points in attack. [199]

2. Steady state.

- (a) Length of steady state. [201]
- (b) Quietest point between attack and decay. [199]

3. Decay.

- (a) Rate/Length of decay. [51], [201]
- (b) Beginning and end of decay points. [199]

4. Amplitude envelope.

- (a) Dispersive or forced continuation shape. That is, quickly rolling off after the attack, or continued excitation. [80], [193], [211], [212], [226]

5. Harmonic form.

- (a) Fundamental amplitude with respect to other harmonics/components [83], [217]
- (b) Spacing / Stretch of harmonics. [51], [52], [215]
- (c) Number of harmonics. [52], [131]
- (d) Balance of harmonic amplitudes. [7], [15], [22], [71], [80] [83], [88], [96], [100], [109], [112], [131], [149], [165], [164], [167], [173], [180], [201], [206], [211], [212], [214], [217]
- (e) Odd/even harmonic amplitude balance. [43], [83], [88], [112], [127]
- (f) Slope of harmonic amplitudes. [52]
- (g) Rate of change of harmonic amplitudes. [164]
- (h) Ratio of sustain and decay times for successive harmonics. [201]

6. Inharmonic form.

- (a) Harmonicity (match to harmonic template). [46], [51], [98], [107], [118], [127] [211], [226]
- (b) Harmonic / inharmonic balance. [80], [212]

7. Patterning of frequency information.

- (a) Proximity of strong partials. [7], [30], [88], [157], [171], [174], [214]
- (b) Distribution of partials. [116]

8. Spectral contour.

- (a) Frequency balance between different regions. [30], [71], [80], [116], [157], [197], [200], [203], [210], [211], [212], [218]
- (b) Positions and amplitudes of formants / resonances. [6], [7], [15], [71], [77], [80], [98], [115], [137], [173], [175], [184], [188], [190], [211], [217], [226].
- (c) Spacing between regions of spectral emphasis. [30]
- (d) Centre/extreme frequency emphasis. [30], [112]
- (e) Significant frequency spread around centroid. [30], [80]
- (f) Spectral slope. [19], [24], [52], [80]
- (g) Centroid frequency. [44], [80], [93], [98], [100], [107], [112], [123], [127], [137], [173], [176]
- (h) Upper cutoff frequency. [24], [22], [77], [127]
- (i) Local gradient of spectrum. [200]
- (j) Long-term average spectrum. [22], [106]
- (k) Onset average spectrum. [106]
- (l) Spectral envelope “parameter”. [168]
- (m) Skewness and kurtosis of spectral form. [80]

9. Noise aspects.

- (a) Spectral element width. Whether tends toward noise or single frequency elements [30], [193], [226]
- (b) Distribution of noise bands in the onset. [200], [226]
- (c) Spacing of pulses in “rustle noise” representation. [51]
- (d) Deterministic/stochastic proportions. [127]
- (e) Spectral density. [107], [106]

10. Temporal evolution.

- (a) Cyclical variations (tremolo, vibrato, spectral vibrato and so on). [7], [22], [30], [51], [171], [172], [174], [175], [199], [226]
- (b) Frequency modulations (rate, depth, onset, non-onset). [25], [80], [106], [150]
- (c) Amplitude modulations (rate, depth, onset, non-onset). [25], [47], [106], [171]
- (d) High-low frequency balance build-up and decay. [172], [226]

- (e) Instability of pitch. [171]
- (f) Shape of harmonic envelopes over time. [109], [136], [137], [138]
- (g) Speed of temporal development. [211], [212]
- (h) Spectral track persistence. [107], [106], [211]

11. Synchrony of partial movements.

- (a) Offsets between rises and decays of harmonics and other strong partials. [22], [25], [47], [51], [71], [80], [107], [106], [118], [138], [165], [171], [172], [212]
- (b) Order of arrival of harmonics and other strong partials. [175]
- (c) Offsets between lower and upper harmonics rise times. [173], [201], [211]
- (d) High harmonic synchronicity in onset. [80], [200]

As is apparent from the above list, there is a considerable range of feature specifications from the vague to the precise. The list attempts to summarise a large range of viewpoints under common headings. As such, in some cases, for example those concerned with harmonic amplitudes, there are more specific details to be found in the references. There is little to be gained in the scope of this research from listing every specific point made by all authors in great detail. Similarly, the vaguest of descriptions have been omitted from the list, such as “spectral changes in the attack”.

It is apparent that some features have a considerable number of references. Considerable numbers of references to attack qualities (such as attack rate), spectral distribution (such as centroid frequency) and spectral “movement” are apparent. However, repeated references do not imply that such features are necessarily fundamental to timbre in all circumstances (see [105], for example). Rather, when researchers have been looking for significance in those areas in particular, it has been found that they commonly have some relevance.

A particular problem in dealing with the list of features is that many of the spectral aspects that are regularly mentioned are often difficult to specify in a manner which allows general extraction of the features from sound stimuli. For example, there is a lack of specific features in the literature concerning transient aspects. Yet, such aspects are regularly mentioned as being important to timbre; they are important in the theory of

specificities, for example. Similarly, “attack quality” is known to be important, but not enough is yet known to specify all the features which are relevant.

From the previous section, it is known that such aspects as the amplitude envelope and spectral envelopes have importance. It is possible to understand the sort of parts which should be extracted from them, from the above list of features. A method is to direct statistical methods to those parts (see [98] for example). This allows more objective and less speculative assessment of the acoustic form, hopefully producing a range of new information, without resorting to guessing the feature parts which are important, or testing each of the large number of features that previous researchers have tested in very different situations. Such a method is used in Chapter 5.

Another problem with trying to categorise the spectral form is that the interactions between components become unclear. In particular it is important to remember that many of the features have a form which varies with time, necessitating further categorisation (see [98] for example). A “single” feature like balance of harmonic amplitudes therefore hides a set of time-varying amplitudes, which will probably have correlated aspects, from which a set of more specific features must be extracted to understand what is really going on within that element of the spectrum.

2.11 Spectral Correlates of Timbral Semantics

“Although musicians possess a very rich vocabulary for describing musical timbre, conventional synthesizers are unable to make use of it.” [52]

Sections 2.9 and 2.10 summarised those parts of the spectral form which have previously been found to contribute strongly to timbre perception in, at least, a moderately general manner. However, it is also interesting to consider how the spectrum relates to verbal descriptions. This section details authors’ attempts in this area. The aim is not to list all descriptions which may be applied to particular instruments, but rather those which were believed to be applicable in a general way by previous researchers.

1. **Acute.** In [190] it was considered that this corresponds to increasing frequency of the second formant.

2. **Bright.** The most common description to occur in the literature, this is often identified as an highly salient/fundamental attribute of timbre. Two spectral effects are normally associated with brightness:

- (a) A low pass filtering type effect. The wider the spectral spread from 0Hz upward, the brighter the tone ([42], [64], [24], [52], [211], [111]).
- (b) Increasing frequency of the spectral centroid leading to a brighter sound ([112], [12], [173], [98], [211]). This effect is also equated to “sharpness” and in [30] to the grave-acute opposition. This may also produce a less pleasant effect than (a) ([13]).

Bright may also be equated to “brilliant”. Ethington and Punch ([52]) describe the brightness effect in slightly different terms. That is, that the number of harmonics increases, the spectral slope increases positively (both a low pass type effect), but also that the density of the harmonics becomes expanded. This latter point might be extended to general partial density (see also [30]). Necessarily, attack quality is linked to rapidity of onset of partials, and thus high frequency components, which directly affects brightness ([77]). Brightness necessarily relates to spectral energy distribution, which is an important characteristic of timbre ([71], [217], [218]).

- 3. **Clang.** The effect of inharmonics can produce a considerable range of timbral change including metallic, bell-like and other percussive/clanging effects ([88]).
- 4. **Clear.** Increasing clarity may arise from decreased number of harmonics, expanded harmonic density and decreased spectral slope ([52]). Jeans ([96]) believed that the 2nd harmonic’s amplitude relates to clarity as it is an octave doubling effect. In [197], clarity positively correlated with high frequency energy, neutrally with mid frequencies, and negatively with low frequency emphasis.
- 5. **Cutting.** Helmholtz ([83]) equated a cutting character to powerful harmonics from the 6th to the 10th, compared to the first upper partials. Similarly, Forrest ([61]) suggested removing lower harmonics makes the sound more cutting. High frequency information can often have obtrusive or annoying effect ([13]), which might be considered brightness taken to the extreme.
- 6. **Dull.** This is generally regarded as the opposite of bright, although Helmholtz [83] also added the qualifier that the effect occurs at low pitches.

7. **Fat.** This term is often used to describe the sound of 1960s/1970s analogue synthesisers. The effect would appear to result from drifting oscillator tuning ([204], [61], [221]) combined with multiple stacked oscillators, tuned to nominally the same frequency but drifting independently. The use of subtractive synthesis in these machines also leads to a large number of partials being present.
8. **Full.** This is attributed to various effects, from having both odd and even harmonics ([153]) to having more low frequency presence.
9. **Grating.** In the attack this may result from inharmonicity ([74]).
10. **Hard.** This effect was equated to increased high frequency content in the attack by Wishart ([226]), who also called it “brittle”. Alternatively it might be due to quick attack of the low harmonics ([74]), or be a correlate of the narrowness of the resonance peaks in the spectral envelope ([190]). Hardness/compactness increases with frequency and intensity, according to [198] (the “density” effect).
11. **Heavy.** This was associated with dominance of energy in the region 150-1200Hz, as opposed to 4800-9600Hz (implying lightness), in [197].
12. **Hollow.** Helmholtz ([83]) said that this results from odd numbered harmonics predominating at the lower harmonics (maybe up to about the 6th). Jeans ([96]) picked out the 3rd harmonic in particular, also relating it to thickness and nasality. White ([221]) similarly considered odd harmonics to cause an hollow sound.
13. **Intense.** Intensity seemed to correlate with the frequency shape of the sensitivity curve of the ear in [197].
14. **Jarring.** This has been equated to strong harmonics all the way to the 16th or 20th and possibly giving a metallic quality when these only die away slowly ([83]).
15. **Lax.** In [190], timbre tended toward a maximally lax position when the first and second formants approached $F_1 \approx 0.6\text{kHz}$, $F_2 \approx 1.4\text{kHz}$ for spoken vowels.
16. **Nasal.** Like hollow, this is as a result of odd harmonics with more strength than other parts of the spectrum ([88]). [83] qualified this with a necessity for energy in the higher partials (above the 6th possibly). The 6th harmonic was picked out in [96] as adding shrillness of nasal quality. But in [112] this seems to be “fullness”. [127] has odd-even harmonics balance.

17. **Open.** This corresponds to increasing first formant frequency according to [190].
18. **Penetrating.** This occurs with inharmonic/dissonant high partial content ([83]), rather than simply high frequency emphasis.
19. **Presence.** Strong components around 2kHz result in presence ([173]). This could relate to the well documented singer's formant effect for voices around the 2.5-3kHz region ([7], [80], [22]), or even the region of maximum sensitivity of the ear at 3150Hz ([198]). Howard and Angus said that the main resonance frequency of the ear is higher than these, at about 4kHz ([88]).
20. **Rich.** This is considered to occur from having the first harmonics ascendent (up to about the 6th possibly) in [83]. From the results of [180], [197] and [167], richness/colour seems to relate to the number of significant harmonics. The 5th harmonic increases richness and an horn-like quality ([96]).
21. **Reedy.** White ([221]) said that a buzzy reediness occurs with a pulse wave, which is characterised by a slow spectral rolloff compared to other elementary waveforms, and both even and odd harmonics present. Howard and Angus ([88]) suggested that emphasis on the 7th harmonic and possibly the 5th as well tend to promote a reedy nature.
22. **Resonant.** This is a similar effect to brightness in that the number of harmonics and slope of the spectrum is increased to heighten resonance. However, the density effect is the opposite - compressed harmonic density ([52]).
23. **Rough.** Roughness seems to indicate effects such as irregularities of vibration ([83], [180], [127]) or mistuning of high harmonics of 1-3% ([174]). In [214], [71] and [173] it is concluded that roughness may result from having more than one successive harmonic (both of significant amplitude) in the same critical band of the ear. Consecutive harmonics above the 6th form intervals of a 2nd and so create roughness ([112]). Jeans ([96]) said that odd harmonics from the 7th upward create dissonant roughness. Plomp ([157]) equated it to a difference of 50-100Hz between partials, depending on the position in the spectrum. Rasch and Plomp ([171]) quite exactly classified differences of less than 20Hz as producing beats (regarded as important in [30]) and between 20Hz and half the critical bandwidth (either side of the partial) as being roughness. This is then qualified with maximum dissonance at about a quarter

Bark interval (1 Bark = a critical bandwidth). These results are often confused by the testing process involving simple tones rather than complex ones and the fact that the critical bandwidth changes with frequency. Roughness is, then, likely to be affected by pitch, spectral structure and musical circumstance. Plomp and Levelt ([158]) also concur with some of these effects. Roughness is also related in some texts to “harshness” and “dissonance”.

24. **Small.** This results from increasing both the first and second formant frequencies ([190]).
25. **Soft.** A lack of harmonics above the first gives a soft tone ([83]), but logically also tends to give a dull one. Softness in this context may also be to do with lower pitches. It might also be considered to lead to a “pure” tone ([167]).
26. **Thin.** Forrest ([61]) said that a lack of fundamental can lead to a thinner sound. A lack of lower frequency components in general can also lead to thinness ([77]). Strong sub-harmonics can have the opposite effect ([220]).
27. **Warm.** Light inharmonicity of the partials nominally at harmonic multiples can produce warmth ([172], [153], [174]) or more natural sound in general ([131]). There is a possibility of roughness, should more than one partial be present in the region of the harmonic, rather than warmth. In [52] warmth resulted from a decreased number of harmonics, compressed harmonic density and decreased spectral slope. Similarly, the results of [167] seemed to indicate a dominance of low frequency energy leading to warmth (in about the first three harmonics). Warm is another term, like fat, applied to the sound of analogue synthesisers and sometimes the effect of vacuum tube distortion. The latter may be due to emphasised second and third harmonics ([108]).

There are problems in utilising the above list of descriptions. As with the spectral features of Section 2.10, the applicability of the results is determined by the timbre space within which they were derived. It is difficult to tell without further experimentation how generally they might be applied. The list attempts to avoid features specific to particular instruments, but does include work by Slawson which is derived from voice sounds ([190]) which has particular interest value, being concerned with formants. The general problems with timbre descriptions, as considered in Subsection 2.8.1, include ambiguity and inability to describe complex differences in a consistent manner. The list must be treated with

scepticism, but aids in building up a picture of the structure of timbre perception from a spectral viewpoint, like the spectral features of Section 2.10.

2.12 Limitations of Information

The information presented in this chapter does not give a singular view of research into timbre. It has attempted to present a wide range of opinions and experimental data to aid in understanding more about what timbre is and its relationship to the time-varying frequency spectrum form. At times information has been presented which deliberately shows how contradictory viewpoints exist in the literature. This section expands on the limitations of the presented information and how different areas should be interpreted.

2.12.1 Timbre Space and Universal Applicability

“One cannot simply apply a process, ‘turn the handle’, and expect to get a perceptually similar transformation with whatever sound source one puts into the process.” [226]

A recurring theme in this chapter has been that there are strong limitations to that which a single study of timbral form can achieve. That is because a study is necessarily constrained by the investigation’s associated timbre space. All except some of the most recent studies necessarily considered very limited regions of timbre space, due to their authors being severely restricted by equipment for data analysis. Earlier studies are also a viewpoint from a more limited knowledge base. As such, if a “rule” is apparent in one instance, with what certainty is it possible to predict that it will be true under different conditions?

A negative answer to this considerable problem is to state that it is impossible to apply results from any studies which did not consider an “adequate” spread of data points. Given that understanding of timbre perception and its relationship to acoustics is at such a primitive stage of development, almost any study could be considered inadequate. Even the 153 sounds which this research considers is a tiny fraction of the total number of perceptibly different sounds in the timbre universe. It is unreasonable to expect to find all the universally applicable facts about timbre at this stage. As with any science, it is

necessary to build up a knowledge base of those aspects of the entity in question which are of importance, and how important, and to find those which are not.

In particular, what is gained from consideration of the general aspects of the spectral form (Section 2.9) is a widely applicable group of information. This provides a solid basis for further consideration of the spectral form. However, the particular features which have been found important by different authors (Section 2.10) contribute *methods* as well as some specifics. It is likely that specific aspects will have more general applicability than specific features, but the features provide understanding as to the *types* of components of the aspects which are likely to be important.

The role of previous studies, therefore, is in providing a means by which it is possible to build up a picture of those parts consistently, often, sometimes and rarely found to be important. To note every single author's findings in extensive detail is not as important as understanding what the major thrust is. It should not be surprising to the researcher, should the applicability of unusual findings be questioned by an experiment. Yet, it should be if those findings do not bare some relationship to the known major areas of interest in the spectral form. Undoubtedly in future the power of computing will facilitate the investigation of much larger (and more densely detailed) timbre spaces, eventually making the worries about applicability less relevant.

2.12.2 Relative Importance of Spectral Components

The *relative* importance of the aspects of the spectral form discussed in Section 2.9 in distinguishing or manipulating among the stimuli of interest is not fixed. Such matters depend crucially, again, on the timbre space. In particular, authors have noted that the relative perceptual strength of the amplitude envelope shape compared to the underlying spectral form is variable ([174]). This may depend on what the envelope shape indicates about the way the sound is produced, or the underlying cues from the spectral features. Pitt and Crowder ([155]) thought that the attack rate was not as important as the frequency content in auditory imagery. But, Pollard and Jansson ([164]) consider that the start of the sound is most important for certain classes of sounds. At a superficial level, this shows that researchers often fail to verify each others' findings. However such differences can be attributed to each researcher using a slightly, or very, different set of

inputs and analysis techniques. For example, Grey ([71]) used very short stimuli which were therefore dominated by temporal features ([80]), placing significant emphasis on the importance of the attack portion as opposed to the steady state.

2.12.3 Composite Form of the Analysis Model

It is apparent that there is a distinct tendency in much of the literature to consider different parts of the spectral form in isolation. This may be a product of experiments investigating simple features acting on simple sounds. It is important to remember that, unless a listener is able to deliberately “hear out” particular properties of sounds, the human aural system assigns meaning to the overall timbral complex, rather than based on single simple cues ([214], [80]). For example, a single harmonic does not define a group of sounds, it is only as a part of a distinctive feature pattern that an identification occurs ([214]). Furthermore, as discussed in Section 2.7, timbral perception is not based on invariant cues, thus a single feature is unlikely to be equally prominent in making a particular distinction between timbre types in all situations. Moreover, there are perceptual interactions between different features ([157]), as well as acoustical ones. Similarly;

“... the processing of individual frequency components is neither independent nor linear. We cannot, therefore, predict the responses to complex sounds by simple summation of the responses to their frequency components.” [150]

Multiple acoustic attributes may contribute to the same perceptual dimensions ([93]). Often, it is only as part of a set of dimensions that a distinction appears to be robust (Chapter 5); single parameters are rarely universally applicable in this regard. As with other decisions regarding how to approach investigation of timbre, expecting distinctions to be governed by singular simple features will colour the results. Therefore, it is important to approach such analysis with an objective viewpoint on how many axes may be necessary and what form they may have.

2.12.4 General Reliability of Previous Research

A major problem in taking authors at their word is sometimes that certain assumptions cloud the reporting process, or remain unstated. For example, “musical” sounds are often assumed to be based on a tonal harmonic form, meaning that harmonics produce dominant timbral characteristics. This is only the case sometimes. Furthermore, it is possible to place a great emphasis on particular elements as exclusive, such as there being a clear divide between that which is noise and that which is a “simple” partial, or similarly something that has continuity and that which does not, or harmonic and an inharmonic form. In terms of the original physical source form, such a divides are nonsensical ([164]). However, in an empirical analysis it is necessary at some point to define where boundaries exist, in order that general conclusions can be drawn regarding the types of elements which contribute to particular timbral perception.

A second part of reliability is that source texts may be treated very loosely, or misinterpreted. Subsequently, such concepts can filter through to become part of timbre theory:

“Certain relationships and concepts . . . have been granted the status of reality through their repeated references in the literature. However, quantitative and qualitative analyses across a variety of sources indicate that this status is not warranted.” [76]

This is an extreme view, and it may be true that some concepts have become so well known that they are often quoted without additional experimentation being carried out. The job of the new researcher must be to carefully analyse how much genuine experimental evidence exists. Often this is indicated by the spectral *features* which have been investigated, rather than the more general *aspects*. This is because some authors are prone to stating such things as “the attack quality is of importance”, without stating what “quality” implies.

2.12.5 Breadth of Consideration

There are some more subtle unstated facts in the literature. These concern the nature of the investigations' search spread. Firstly, it is only possible to consider a limited number of features within each spectral aspect area within a single investigation. As such, although a feature may be found to be particularly important in a distinction of interest, or possibly two or more features, such features are not necessarily the best for the task. That is, the feature(s) is best out of those considered, but more effective features might have been derived, within the timbre space of interest, which would have been *more* effective. For example, attack time might have been measured, but attack time : stimulus length ratio may have been more effective in the distinction of interest. As such, it is always an healthy tactic to expand on previous research techniques, rather than limit the current experiment to well worn features, while the true nature of timbre is still hazy. Similarly, *expectation* of a feature having salience is an unsound tactic, because so many features of the spectral form have some kind of relationship to perceived timbre. The question is, whether the particular well known feature is most important in the particular circumstance. That is where objective analysis, of the type described in Chapter 5, becomes particularly vital.

Research can be coloured, therefore, by choosing a particular way of viewing the space through a feature set. Similarly, the spread of the features over the structure of the timbre space is of importance. For example, in a hierarchical form, do the features adequately describe both minor and major differences, as described by the levels of the structure, in adequate detail? In a form based on specificities, do they form a set adequate to describe both the major characteristics and nuances on top of those major parts?

2.12.6 Quantification

A particularly apparent feature of the literature is a lack of quantitative information. This is sometimes because the results are based on observation of characteristics (for example, matching spectrogram features by eye), or correlations (for example, description X being generally correlated with spectral feature A). Even when results are produced from a numerical process, the relative magnitude of different effects is rarely documented. This indicates the difficulties of both quantifying differences and conveying results in a interpretable form. Often, it is the fact that an effect exists in a particular form, within

the confines of the timbre space under consideration, which is of interest. As every timbre space is different, exact quantification can be meaningless.

2.12.7 Role of Timbral Semantics

Trying to describe timbral effects is a particularly difficult problem, but is also one which allows a sensation to be conveyed to another person in terms which are not only degrees of difference. In this way it is much like wine tasting. Also, verbal ratings can lead to structured descriptions which can be correlated with acoustical differences. Subsection 2.8.1 described previous attempts to understand different timbral descriptions, and the associated structure. Section 2.11 described some of the postulated spectral correlates of timbral descriptors. The range of perceptible timbral nuance far exceeds linguistic ability, however ([171]). Also, the twin problems of ambiguity and limited quantity of descriptions apply. The fine nuances of differences between, say, a violin and viola, or even two violins, playing the same note are indescribable in a consistent and globally applicable manner. Judging degrees of (dis)similarity have proven more effective and consistent.

2.12.8 Treatment of Individual Instruments

This chapter has not covered specific details of specific instruments. The primary reason is that this research develops theories concerning timbre and timbre perception from the spectral viewpoint, rather than that of the physical description of instruments. As such, interest lies in those aspects of the spectrum found to be important to timbre perception, in a general manner. Discussion of specific instruments' physical, acoustical and timbral forms would necessarily be cursory in a text of this size. Extensive details can be found in such texts as [22] and [60].

2.12.9 Consideration of Higher Level Concepts

The previous sections have been mainly concerned with "elemental" forms. There is a good reason for this, which is that high level constructs are out of the range of current understanding. The processes by which research aims to progress is by building up layers of lower level features and attempting to correlate them with higher level perceptual

structures. However, some authors have attempted to analyse timbral description at a very high level ([131] for example). Such descriptions include “natural”, “boring”, “complex”, “pleasant” and so on. These border on the line between timbral aspects that can be equated to groups of definable parameters and personal preferences or vast comparisons between groups of sounds. It is always tempting to abstract as far as possible when considering timbre, as it is naturally a subject concerned with high-level description. As will be seen in Chapter 3, humans are highly capable of making high level perceptual judgements concerning timbre in a organised manner, but the cohesiveness of judgements between subjects improves between judging at a very high level such as how “synthetic” a sound is, to a less high level, such as how “woodwind”-like. It is certainly desirable to be able to establish the sorts of effects which constitute very high level characteristics, such as “naturalness”. But, the higher the level, the more likely it is that a single umbrella characteristic will in fact represent not only complex combinations of features, but also take a number of different forms and be subject to considerable variation with timbre space and listener.

2.13 Conclusions

This chapter is novel in the following ways:

1. The chapter is the most comprehensive overview of the sound timbre literature and concepts to date.
2. The analysis of the components defining the concept of timbre has not been previously attempted.
3. The in-depth consideration of the definition, factors affecting, dimensionality and structure of timbre space is more wide-ranging than previously considered.
4. The review of the spectral aspects and features relating to timbre in the literature is the most comprehensive to date.

CHAPTER 3

Perceptual Study

“The concept of the relative distances of musical tones, in terms of perceptual contrast or dissimilarity is not a foreign one in the listening experience of the musician.” [71]

3.1 Introduction

This chapter discusses the nature of timbre perception and a study concerning the perceived similarities between the group of 153 sounds used in this research work (Appendix A), as judged by 14 participants of varying musical background. The contents are as follows:

1. The reasons for the perceptual study.

2. An overview of the mechanisms of timbre perception.
3. The perceptual study technique employed.
4. A statistical analysis of the result data and its implications.

3.2 The Reasons for the Perceptual Study

“Due to its overwhelming complexity, timbre perception is a poorly understood subject.” [71]

As late as 1975, some scientists still doubted that psychological magnitudes of sensations could be usefully assessed ([198]). As Stevens pointed out;

“Nothing stops research more effectively than the belief that a kind of measurement is impossible.” [198]

There is now considerable evidence to suggest that the magnitude of timbral change can be assessed, and related to acoustical measurements.

“The typical observer’s capacity for magnitude matching is generally far greater than some experimenters have seemed willing to believe.” [198]

As discussed in Chapter 2, it is possible to take different viewpoints on the investigation of timbre, but the two main directions of consideration are from either a perceptual, or an acoustical viewpoint. If research only considers the acoustical attributes of sounds, then there is a danger of failing to understand the relationships between sounds as perceived by the human brain. For example, it might be *assumed* that strong axes of timbre are those which can clearly distinguish sound group X from group Y, when in perceptual terms X and Y may share many characteristics. From the opposite perspective, only considering perceptual relationships fails to indicate those parts of acoustical models which are of relevance, and thus how analysis, modifications and synthesis might progress.

This perceptual study is designed to establish the following concepts:

1. To show that timbre space and timbral relationships are not only an engineering model, but also a psychoacoustic one. The concepts of timbre spaces introduced in Chapter 2 represent a very convenient way of describing sound quality forms and their relationships. However, such concepts are also meaningful with respect to how the relationships between sounds are *perceived*. It is fundamental to timbre research that the acoustical and mental timbre/instrument spaces can be related.
2. To show that knowledge of timbral relationships is partly independent of musical training, and is a natural part of perception. That is, at a basic level, all people have the inbuilt ability to distinguish sound qualities in a logical manner, and discriminate fine nuances of timbral change. That is not to say that practice in listening to the differences between sounds is not of relevance, as it may enable better judgement. However, prior knowledge might also cause a listener to fail to recognise important aspects of the sound through familiarity (Subsection 3.3.6).
3. To show that relationships between timbres are perceived in a logical, structured manner. This extends the last point, in that it is apparent that not only is timbre an inherent part of the human experience, but also that there is a cohesive structure involved in the perceived relationships between sounds. That is, sounds are entities which are not generally perceived as being isolated.
4. To show that perception of timbre is continuous, rather than categorical. Sounds can be recognised as belonging to a continuum of change within the mental structure of timbre, and this is an important part of the timbre space form. It allows instruments to be considered as describing a region of timbre space, dependent on playing style and so forth. Thus a single instance of sound represents a point within that continuous region.
5. To show that instruments' timbre spaces are not neatly separated in perception, but have intersecting characteristics. Different instruments are not mutually exclusive in their acoustical or mental representations. As such, a single instance of sound might not imply a single source, but rather a number of instances might be required to make a correct categorisation judgement.
6. To show that the mental timbre space described by the range of sonic information in a set of template sounds, can be successfully used to match the features of stimuli

which are not part of the template. This will be explained in greater depth later in the chapter.

3.3 Mechanisms of Timbre Perception

“It has also been demonstrated that the hearer perceives a relation of distance between timbres, which he organises in a space where each sound occupies a place by the function of its colour.” [91]

The science of perception is complex and only partially understood. This section provides a brief summary of the state of knowledge concerning timbre perception, and so the background to the experimental details of Section 3.4 which follows.

3.3.1 Structure of Auditory Information in the Brain

The processes of auditory perception would appear to include the following aspects ([5]):

1. Encoding of low-level features such as frequency relationships, gradients, intensity and movement information.
2. Encoding of high level features of sound stimuli.
3. Reduction of information quantity into a concise form.
4. Multiple representations of the external World.

There are a number of key features in the above list. The first two items hypothesise a staged approach to dealing with information. Rather than auditory perception being based on a single-stage analysis of the raw filtered information extracted in the peripheral hearing system, it seems likely that the mind builds up a picture of the sound based on combinations of features ([164], [200]). The third point indicates an analysis based on reduction to particular salient features, possibly for classification purposes (this is discussed later in the chapter and in Chapter 4). The final point is particularly interesting in that it indicates that perception is not governed by fixed rules, but may adapt to choose an appropriate viewpoint (also discussed later).

At a peripheral level partial encoding of acoustic dimensions occurs, and in the central auditory system, peripheral properties are integrated into more abstract representations ([154]). The complexity involved means that timbre is, fundamentally, a central processing phenomenon ([122]). Inevitably, this complicates the situation as timbre research is not able to search for fixed mechanisms, but rather is attempting to model processes that can adapt depending upon the sonic context, prior knowledge, where attention is directed, and so on. However, the fact that models have been constructed based on acoustical features which result in a perceptually logical structure indicates that, while the use of information is adaptable, it is also based on consistent types of analysis ([201]).

The processing which occurs in the auditory cortex has relationships with other parts of human sensory perception. In particular, vision and audio are believed to be processed through some common stages in the cortex ([154]). Vision is often used to augment aural information when someone is speaking, for example, as the mind is gathering as much related information on the subject as possible ([121]). This also indicates the potential for perceptual interactions.

“... the similarity patterns generated by the stores of visual and auditory information may be used to index similar experiences. They can therefore be used to regulate access to related experiences.” [5]

Such links between the senses include some evidence of synæsthesia ([50]). It is certainly true that cross-modality matching is possible between the senses ([198]). Also, many of the adjectival descriptions of sound qualities relate to other senses and are understandable in those terms (Section 2.11). An occasional analogy used with sonic entities is with physical objects:

“An object is an individual, distinguishable entity. We perceive objects. We manipulate them with our hands and we manipulate their representations in our minds ... We are able to manipulate the representations of several objects at the same time, and we are able to imagine interactions between them just as if the objects themselves were present.” [5]

Although the visual and tactile senses are not the same in composition as hearing, it seems plausible that if it is possible to conceive of the form, interactions and manipulation of

solid objects in the mind, then similar processes may exist for sonic perception. The result of this is that it is reasonable to ask subjects in perceptual testing of timbre to relate sound qualities in the way that it is possible to relate visual, or tactile experiences. These reinforce the timbre space concepts because timbre space is to do with relationships through similarity (or lack of it). It is already known that the effects of timbral imagery can be empirically demonstrated ([36]). The concepts of moving through timbre space are analogous to differences in the quality of other senses. Such relationship properties make sense in all senses from gross similarities, to particular aspects and nuances.

3.3.2 Contextual Perception and Directed Attention

As timbre perception is determined significantly by central processing, the context in which sound qualities are experienced can have a significant effect on perception ([68], [80], [122], [164]). Such context can either be subconscious (directed by external events) or conscious (a decision to focus perception on particular aspects of sound). There is a trade-off between these two aspects. If the subconscious cues are too strong, it is impossible to focus on other aspects, for example.

Presentation is important in perceptual testing, as it has a strong influence on which subconscious assessments in the cortex are stimulated. In this research, as discussed in Chapter 2, the focus is on “isolated” individual instances of sound. Other research has included consideration of sequences of sound instances, simultaneous sounds, and transformations and transitions between sound types ([73]). In all of those forms, there is a strong temporal proximity aspect. The importance of time in perception means that when testing subjects there is always the effect of those sounds which have recently been presented. Also the time displacement between the subject being played a comparison stimulus (“adaptor”, or standard against which the present sound is being judged) and the test stimulus can have an effect on the result. The distribution of the sound stimuli over the subjective scale range is important as well ([198], [44]). If stimuli are concentrated in particular regions, then perceiving relative similarity between those groups is obscured by the clear similarity within the groups. Also, the perceptual range covered by the stimulus set affects ability to perceive different nuances of sound quality ([123], [13]). If the sounds cover a large distance in timbre space, it is much harder to rate fine differences in a consistent manner, compared to a smaller perceptual range.

As regards conscious perception, it is known that different types of listening lead to different responses. Gaver ([67]) divides audition into “everyday” and “musical” listening. The former is the most common sort of listening and relates to perceptually identifying the event that caused a sound. This occurs through perceiving the entire sound complex (which might also be called “holistic” perception). Musical listening concerns perception of the component properties of the sound (“analytic” perception, also known as “hearing out” the component parts). This is the mode of listening used in most sonic perception experiments, including the one detailed here.

Attention may be focused at a number of levels below that of the overall sound. At the lowest level, individual partials can be perceived under some circumstances, such as the first 5-7 harmonics ([134]). At an higher level, groups of spectral elements may be perceived which form complex perceptual systems such as brightness, harshness, and openness. Schouten points out, however, that;

“Acute observers may bring some of these elements to conscious perception, like intonation patterns, onsets, harshness, etc, even so, minute differences may remain unobservable in terms of their auditory quality and yet be highly distinctive in terms of recognising one out of a multitude of potential sound sources.” [179]

A potential problem in experiments where a subject is asked to hear out particular aspects of sound sources, is that it is difficult to know if that person is focusing from the big picture to smaller details, or whether perception has been radically altered and no longer represents “normal” audition. An analogy in the visual sense could be the difference between being asked to focus on finer details of a landscape, which are perceived as part of the whole when appreciating the entire scene, and attempting to see the hidden picture in a random dot stereogram, which in no way represents part of the normal viewing experience. When focusing on details, the researcher must be assured that perception is not being abnormally distorted.

Because perceptual analysis is dependent not only upon the resolving capabilities of the peripheral auditory system, but also fundamentally upon central processing abilities, it is necessary to cause attention to be directed to the parts of the sound of interest ([110]). In similarity testing, subjects are asked to direct attention to the details of sounds - to make

conscious rather than sub-conscious assessments. The aim of this is to force the auditory system into assessment of the underlying timbral properties through musical/analytic listening, rather than making sub-conscious categorisation decisions as often is the case in “normal” listening.

“... constrained categorization, comparison, and rating tasks can be used to induce listeners to focus on information for particular attributes at a desired level of detail.” [66]

In particular, the desire is to make the subject being tested utilise a considerable range of attributes, which might be a different set or have different emphasis to those used in everyday listening ([198]). For example, if judging similarity between tones, in everyday listening the loudness, pitch and duration of the sound and classificatory perception might dominate. When quantifying timbral differences, such effects must somehow be nullified by appropriately directing attention (see Subsection 3.4.2).

3.3.3 Continuous and Classificatory Perception

From an evolutionary perspective, the human aural system is designed to facilitate sound object identification ([179], [67], [66]). This implies categorical judgement of the stimuli at some level. This does *not* automatically imply that perception is generally a classificatory phenomenon, though ([121]). This is underlined by the fact that humans do not only perceive general source characteristics (“it is a violin”, “the word being spoken is phantasmatron”), but have an high aptitude for recognising timbral nuance (such as identifying a particular speaker, or even a particular make of violin). This is relating sounds to perceived sources ([224]), but also at multiple levels of detail.

But aural perception goes beyond the simply classificatory, as it is possible to perceive sonic ambiguity ([226]), indicating ability to relate a sound to a number of sources along continuous dimensions. Thus mutually exclusive classification is inappropriate. Where classification is necessary, the mind must make a judgement based on greatest likelihood criteria ([202]). That is, the acoustical attributes would tend to indicate a certain solution, based on continuous information ([121]). Furthermore, some sounds cannot be related to a physical source due to their acoustical composition, and sounds can be produced which

will not fit into a category of sound types previously heard. However, even in these cases, aspects of the sounds can be related to others which have been heard before to fit them into the mental structure of timbral understanding.

It is apparent that there is more than one process at work in perceiving sound timbre. The ultimate aim may be to classify the sound, as far as the auditory cortex is concerned, but that would appear to be the final stages of the process. Pattern recognition is a mapping from a measurement space, to a feature space, and then to a category space ([202]); equivalent to spectral space, timbral space and high level classificatory space. As discussed in Subsection 3.3.2, it is possible to direct attention to particular aspects of the feature space, rather than the final classification. In perceptual testing, it is desirable to explore the mechanisms of perception, by asking subjects to judge similarity based on the underlying sound qualities rather than the overall classificatory picture. Identification and discrimination seem to occur in different cortical domains ([80]). In particular, being able to discriminate two stimuli does not mean that they are not confusable. Confusion in classification occurs over ranges of discrimination.

Therefore, it would seem that just as the structure of auditory information in the brain is built up in a number of levels, it seems possible to perceive that information in a similar manner. Although there is no evidence of direct appreciation of the very lowest level information from the peripheral hearing system, it seems likely that there are layers of continuous perception, coupled to those of categorical perception. The ability to perceive a continuum of timbre has been demonstrated. Without a continuum, sonic morphing would not be possible. Interestingly, some authors ([71]) report cases where it is possible to interpolate from one sound to another, but there is a region of abrupt, if continuous, change. This indicates a combination of the continuous and categorical aspects at work. These combine in the timbre space representation; essentially a continuous perceptual space, but with regions which are associated with particular sources.

3.3.4 Processes of Judging Similarity

It has already been discussed how there seem to be a number of layers of perception in the auditory cortex. However, that does not explain how they relate to judgements of similarity, or source identification (see Subsection 3.3.5). It would seem that there are two

general ways of considering perceptual similarity ([205], [186]):

1. Dimensional similarity. This is where qualities are compared along continuous and homogeneous perceptual dimensions. This type of similarity apparently relates to perception of simple sensory or metric stimuli, such as colours or geometric forms.
2. Symbolic similarity. This is where a number of qualitative features combine to allow classification judgement. This type of similarity may be used for perceiving complex stimuli, such as faces. Symbolic forms cope particularly well with context factors.

These types of similarity judgement relate to the different levels of auditory processing. Symbolic similarity is the more complicated because it is not simply a geometric representation, but takes the continuous inputs and creates particular solutions. Two methods of representing symbolic similarity are;

1. Feature approach. This postulates that similarity is an increasing function of the number of properties that the two stimuli have in common ([205]).
2. Class approach. This suggests that similarity is relative to how many stimulus items there are in each class, and how many in the stimulus universe ([186]).

The reason why this discussion is of interest is that it can help predict how judgements of similarity between sounds will progress. Timbral forms are complex stimuli, and at the highest level are judged in a symbolic manner. However, in the judgement process, attention can be directed to particular elements; the subject can be asked to concentrate on the acoustic differences (such as in this research) rather than on the overall classification level. Considering the most salient dimensions of sound, it is apparent that pitch, loudness, duration, and possibly brightness and attack rate, are judged in a more dimensional than symbolic manner.

Therefore, there is a combination of effects involved. If the subject is being asked to concentrate on the timbral differences, this represents focusing on dimensional forms and thus a continuous solution space. However, the process involves the subconscious symbolic similarities as well. This means that the number of items and spread of sonic range in the stimulus universe will be important ([186]). The bigger the range of stimuli, the more

similarity is apparent overall, and so smaller differences will go unnoticed. Secondly, the stronger the definition of the class to which items are being compared, the more precise the matching effect. If the adaptor (comparison) class is very vague, then a large number of stimuli will appear to fit in the class, as well as there being larger room for interpretation.

In this research, there is a relatively large number of stimuli, covering a considerable range of timbral forms. Thus in similarity judgements, the finest differences will be less apparent. Also, the types of similarity tests (see later in the chapter) have a range of precision involved in their class definitions, and this will lead to a range of vagueness in the judgement of similarity.

3.3.5 Processes of Source Identification

For the human aural system to be able to gain extra information concerning what is happening around the body, and to choose an appropriate course of action, implies relating acoustic input to perceived source. The way that the human auditory system reacts to electroacoustic and “unnatural” sounds in general is difficult to predict, as the perceived gestures will often fail to relate to a known source type. Unknown sounds can be comprehended in terms of more familiar sounds, which form a reference framework ([157], [97], [193]). Handel ([80]) suggests that timbre perception is based both on recognising production invariances, but also linking the acoustical features to objects through experience in different contexts.

As explained in Subsection 3.3.3, it is possible to perceive complex nuances in sounds which cannot always be categorised by “natural gestures”. This suggests that the cortex uses the components of sounds previously experienced to piece together an understanding of new sounds ([5]). Thus, in a classification experiment, “is the sound woodwind-like” is liable to illicit a probabilistic estimate of group membership based on the common features of the stimulus to the woodwind template, rather than generating a response based on perfect match to a category that is mutually exclusive from any other. Again, this indicates classification may be a secondary, higher level of understanding which develops from lower level similarities. Further evidence is provided by Palmer ([150]), who indicates that perception in mammals does not seem to be orientated along strictly functional dimensions. The human hearing system has not been specialised in order to allow direct

understanding of distance, like a bat, for example. Humans seem to perceive based on a picture composed of feature dimensions; adapting to understand many sorts of sound.

It is hard to explain how subjects can perceive object constancy across a wide variation in the acoustical properties of the stimulus ([154]). Handel suggests that:

“First, acoustic properties adhere to objects; the properties belong to and at the same time characterize the source. Second, these properties evolve over time. The changes typically are slow, continuous and regular” [80]

Identification relates to the dimensionality and perceptual structure concepts outlined in Section 2.7. If perception is governed by a “plain” multiple-dimension form, then combinations of particular regions of feature axes might indicate an high probability of a particular instrument classification. With a small set of fundamental axes and instrument specificities form, a region of the important axes and a group of the unique features would be important. With an hierarchical system, the region of interest is refined through a series of stages and the lowest level(s) represent the variations of a particular instrument type and its nuances. All perceptual structure models must be able to cope with disparate acoustic conditions leading to similar perceived timbre, and sometimes the opposite.

It is also worth remembering that in everyday circumstances, the auditory system is usually presented with a number of related stimuli, which help map out an area of timbre space, thus aiding in object identification ([80]). In experiments such as that detailed in this chapter, the subject is presented with only a single instance of sound. Such limited information can lead to confusability (see Subsection 3.5.3). However, it is also a useful experimental tool as it can aid in forcing the subjects to use continuous perception of the lower level features rather than categorical perception based on a larger body of information. Through a carefully chosen lack of data, conscious perception may be more analytical and so focused on the features with which this research is concerned.

3.3.6 Effects of Prior Knowledge

Timbral perception and its classificatory partner, source and event identification, are not cognitive processes which must be learned deliberately ([66]). Knowledge concerning

timbral form is a natural part of understanding the World. The question is how much experience is used to augment that natural understanding. Such effects depend upon the structure of aural processing. Experience will produce more complex effects depending upon the amount of interaction that occurs within the structure of perception. It might be that the cognitive processes are segmented into independent regions of peripheral processing, feature extraction and source/event categorisation. Or, considerable interactions might occur to aid in building a more cohesive picture. It has already been mentioned that it is very likely that different levels of detail are probably used under different conditions, which implies higher level processes affecting lower level analysis.

In the literature some researchers have tried to estimate the effect of experience on perceptual judgements. This in general involves a comparison of musicians and non-musicians, or the researchers' results against those without prior knowledge of the test. Results from different groups offer conflicting views ([105], [43]). Moreover, it is very hard to quantify each participant's lifetime's experiences of sound qualities.

On the one hand, repeated exposure to musical instruments and listening tasks is likely to improve a subject's ability to hear subtle nuances of sound quality, which non-musicians may fail to appreciate. However, repeated exposure may improve classification abilities, meaning that the subject has a tendency to classify, rather than to carefully analyse the feature contents with respect to other sounds (to analyse the sound as if it were a new experience). This implies that a musician, or the researcher who devised the experiment, may find a greater perceptual dissimilarity between the sounds. This is known to be true with vocal sounds ([192]). Another viewpoint is that moderate levels of knowledge may lead to one-to-many mappings, where a feature set may indicate not one, but several sounds or events, blurring rather than clarifying the picture ([66]). Furthermore it is difficult to quantify how different subjects may be able to understand and perceive abstract concepts such as those which are often part of timbral perceptual testing ([123]).

In addition, there is a necessary trade-off between the complexity of cognitive processing (and therefore the amount of prior knowledge involved), and the quantity of information present in a stimulus ([66]). Experience may, then, have a weaker role if more identification cues are present in the stimuli being considered. Tests for systematic differences between different groups of participants performing judgement experiments have been inconclusive. Given the complexity of the situation as outlined above, and the

principle that everyone has the ability to perceive timbral similarities (which is one of the concepts being considered in this chapter), such a result is not surprising.

3.4 Perceptual Study Technique

The perceptual study in this research is designed to facilitate the establishment of the principles outlined in Section 3.2. It is based on judgements of perceptual similarity, as opposed to verbal/semantic rating scales, to avoid the known problems of methods based on description (Section 2.8). It has the following major characteristics:

1. It is based on template similarity judgements. That is, comparing sounds to a *set* of adaptors (comparison stimuli) which describe the timbre space region of interest.
2. It is composed of 6 “tests”, each of which consists of rating the similarity of the 153 sounds outlined in Appendix A to a template of 4 sounds representing a timbre family.
3. The 6 timbre families are String, Woodwind, Brass, Hammered Tonal, Percussive and Synthetic/Test-Tone sounds. These represent a (loosely) traditional set of groups for investigating (dis)similarities between sounds (see Subsection 3.4.2).
4. Each sound from the set of 153 is rated on a four-point scale of template-like, template-ish, hardly like the template, and not like the template.
5. 14 participants were involved in generating the results, including the author, musicians, persons involved in sound research, and also non-musicians.

The study displays some departures from the previous studies described in Section 2.8.2. Firstly, it considers a larger number of sounds and spread of timbre types than has been previously attempted. This should aid in achieving more general understanding. Secondly, the study uses matching based on similarity to a template of four sounds, rather than pairwise or triadic matches. This is a deliberate ploy to achieve better understanding of timbral subspaces and avoid problems of non-timbral equalisation, as described in Subsection 3.4.2. Finally, due to the template similarity technique, it becomes possible to consider what timbral families actually mean in perceptual terms.

Note that this study concerns an examination of discrimination, not the identification of stimuli. As described previously in this chapter, a major aim of the study technique is to cause the attention of the subjects (participants) to be directed to the lower levels of detail, rather than concentrating on classification and “everyday” listening. That is, such that the participants use all the available information in the stimuli, rather than concentrating on particular aspects. This is to allow examination of the nature of the features which underpin the perceptual comparison of timbral qualities. In particular, the relationships between the stimuli and groups of similar sounds is of great interest.

3.4.1 Questionnaire Format

Figure 3.1 shows the first page of instructions issued to participants in the study. It is worth noting that:

1. The template for each test consisted of 4 sounds to provide a clear description of the range of qualities which describe the timbral subspace to be compared.
2. The template was not presented before every sound to be rated, due to time limitations. However, the participants were allowed to hear the template at any point in the test (and often heard it at least once more). Also, the test was not started until the author and subject were satisfied that the task and template were clear to the participant.
3. The gap between each rating item was chosen to be 3 seconds by the author, which in preliminary testing appeared to be long enough to make a decision, but not long enough to allow complex higher-level processes to cloud the judgement. The nature of a template match as opposed to a one-to-one match implies a more complex timbre subspace (rather than timbre distance) matching process which requires more thought. Thus a shorter gap was inappropriate.

Figure 3.2 shows the top parts of the response forms for the 6 tests. It is notable that:

1. Specific textual instructions are provided for each test, to augment the acoustic information. This was explained verbally if necessary, with minimum recourse to specific examples.

2. The rating scale has 4 points. This is an even number to prevent middle judgements which are known to present result-bias, more than 2 which is a yes/no choice, and fewer than 6 which seemed too fine a scale to maintain consistency over such a large stimulus set.
3. The response boxes are labelled to aid understanding and consistent response, rather than rely on the participants' consistency of application of an unlabelled or numerical four-point scale.

There is a fine balance to be struck between explaining the task adequately such that the subjects approach the test with the same frame of reference and listening attitude, and over-explaining and adversely biasing the test. It is apparent from the notes accompanying the response sheets that the author was particularly worried about the subjects slipping into a "traditional" mode of listening and classifying based on high level knowledge rather than acoustic features. The whole study style is to direct attention to the elements of interest.

Further points of note concerning the application of the testing procedure are as follows:

1. The participants performed the tests individually, rather than in groups.
2. The participants in general took either one or two tests at a time, lasting respectively approximately 15 or 30 minutes, to prevent fatigue affecting the results. Nobody participated in more than one testing session per day.
3. The order of presentation of the 153 stimuli was determined by a pseudo-random number generator to reorder the sounds from their grouped arrangement of Appendix A. The presentation order was the same for all subjects and all tests.
4. The stimuli samples were played by a computer, to allow the test to be paused, to allow re-evaluation of sounds if necessary, and to play the template sounds again, with ease.
5. The stimuli were presented on headphones.

Questionnaire to Establish Perceptual Similarities for Timbral Families

Version 1.0, David P. Creasey 7/1/97

Estimated duration = 1 hour 30 mins in total

Instructions

This questionnaire aims to establish perceptual similarity ratings between a group of 153 sounds, and 6 timbre families. The families are as follows:

- (1) String type sounds
- (2) Woodwind type sounds
- (3) Brass type sounds
- (4) Hammered tonal sounds (e.g. piano, marimba)
- (5) Percussive sounds
- (6) Synthetic / test-tone sounds

Each family will be considered separately:

- (1) Four sounds will be presented which are typical examples of the sound group of interest to establish the template timbre type (in conjunction with written notes).
- (2) The 153 sounds will be presented, with a short gap between each for the participant to rate the sound's similarity to the group of interest in the boxes provided.

Notes:

- (1) The aim is not to group the sounds exactly in the traditional orchestral sets, but to classify with respect to the timbre type. For example, a saxophone may traditionally be described as woodwind, but some examples are as brassy as they are woody.
- (2) Only mark one box per sound.
- (3) The sounds have an amount of background noise, which should be ignored in judgements.
- (4) The typical/template sounds can be repeated at any time during the test, if required by the participant. Also, the process can be halted temporarily, and individual sounds can be repeated if required.

Many thanks for taking part,

DPC.

Figure 3.1: Instructions for Perceptual Study

Timbre Group 1 : String
Notes: The typical sound categorised as string-like is anything stringy, no matter what the playing style. Thus, if a sound is plucked or bowed or struck, the aim is still to find its string-like character as typified by the template examples.

| | String-like | String-ish | Hardly stringy | Not stringy | | String-like | String-ish | Hardly stringy | Not stringy |
|---|--------------------------|--------------------------|--------------------------|--------------------------|----|--------------------------|--------------------------|--------------------------|--------------------------|
| 1 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 78 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 2 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 79 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 3 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 80 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 4 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 81 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 5 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 82 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 6 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 83 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 7 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 84 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

Timbre Group 2 : Woodwind
Notes: Again, playing style is of no consequence. But, note sounds which are partly wood-ish and partly brass-ish. Also pay particular attention to any breathiness and the beginning of the sound.

| | Woodwind-like | Woodwind-ish | Hardly woodwindy | Not woodwindy | | Woodwind-like | Woodwind-ish | Hardly woodwindy | Not woodwindy |
|---|--------------------------|--------------------------|--------------------------|--------------------------|----|--------------------------|--------------------------|--------------------------|--------------------------|
| 1 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 78 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 2 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 79 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 3 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 80 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 4 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 81 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 5 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 82 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 6 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 83 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 7 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 84 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

Timbre Group 3 : Brass
Notes: Similar notes to woodwind apply.

| | Brass-like | Brass-ish | Hardly brassy | Not brassy | | Brass-like | Brass-ish | Hardly brassy | Not brassy |
|---|--------------------------|--------------------------|--------------------------|--------------------------|----|--------------------------|--------------------------|--------------------------|--------------------------|
| 1 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 78 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 2 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 79 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 3 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 80 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 4 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 81 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 5 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 82 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 6 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 83 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 7 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 84 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

Timbre Group 4 : Hammered Tonal
Notes: Here the aim is to find a playing style / amplitude characteristic in particular (a hammered attack with a gentle tail), but also linked to the emphasis on lack of noise. Thus a struck cymbal is not as hammered tonal as a piano.

| | HT-like | HT-ish | Hardly HT | Not HT | | HT-like | HT-ish | Hardly HT | Not HT |
|---|--------------------------|--------------------------|--------------------------|--------------------------|----|--------------------------|--------------------------|--------------------------|--------------------------|
| 1 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 78 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 2 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 79 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 3 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 80 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 4 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 81 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 5 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 82 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 6 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 83 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 7 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 84 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

Figure 3.2: Top Parts of Response Forms for Perceptual Study

| Timbre Group 5 : Percussive | | | | | | | | | |
|---|--------------------------|--------------------------|--------------------------|--------------------------|----|--------------------------|--------------------------|--------------------------|--------------------------|
| <i>Notes: Playing style / amplitude characteristic is the important factor here. We are looking for a sharp attack with a quick rolloff, independent of underlying tone colour. That is, not simply a short sound, but a sound where the attack and initial decay are brief and the rolloff overall is quick.</i> | | | | | | | | | |
| | Percussive | Percussive-ish | Hardly percussive | Not percussive | | Percussive | Percussive-ish | Hardly percussive | Not percussive |
| 1 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 78 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 2 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 79 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 3 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 80 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 4 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 81 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 5 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 82 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 6 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 83 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 7 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 84 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

| Timbre Group 6 : Synthetic / Test Tone | | | | | | | | | |
|--|--------------------------|--------------------------|--------------------------|--------------------------|----|--------------------------|--------------------------|--------------------------|--------------------------|
| <i>Notes: Playing style / amplitude curve are unimportant. The aim is to segregate those sounds which appear to be from a 'natural' source and those which are synthetic. Note also that some sounds in the set are from a synthesiser wavetable and may be regarded as synthetic-ish.</i> | | | | | | | | | |
| | Synthetic | Synthetic-ish | Hardly synthetic | Not synthetic | | Synthetic | Synthetic-ish | Hardly synthetic | Not synthetic |
| 1 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 78 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 2 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 79 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 3 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 80 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 4 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 81 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 5 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 82 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 6 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 83 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 7 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | 84 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

Figure 3.2(cont.): Top Parts of Response Forms for Perceptual Study

3.4.2 Templates Employed

The template technique avoids some of the major problems associated with similarity judgements in previous experiments, by asking the subjects to fit stimuli to a perceived instrument space, rather than to match sounds one-to-one. The templates used for comparison in this perceptual study were partly designed to represent the broad traditional categories ([32]) of Western orchestral instruments (Strings, Woodwind, Brass) with the omission of the vague and wide ranging "Percussion" type. Rather, the more complex concepts of Hammered Tonal, Percussive and Synthetic are used to attempt to delineate more accurately the perceptual boundaries of the sound types, and thus achieve more meaningful results. As is pointed out in [22], treating these categories as mutually exclusive, though, is a mistake:

"In the orchestra it is normal to divide the instruments into four classes, strings, woodwind, brass and percussion. This classification is convenient

because it relates directly to the sound quality of the instruments and also the positions at which they are situated within the orchestra. However, it is not based on acoustical principles so it will not help us with a scientific understanding of how the different instruments work. For example it does not take account of the fact that in both brass and woodwind the sound is generated by exciting a column of air within a tube.” [22]

This means that the traditional segmentation of instrument types is artificial. This is an advantage in this perceptual experiment as it enables the examination of how subjects are dealing with the acoustical information. If subjects are concentrating on the traditional, high level classificatory perceptual level, then it would be expected that the subjects would segment the different groups of sounds in a traditional manner. For example, a classificatory response would be that a string sound bears no fundamental similarity to a woodwind sound, by traditional definition. The desired effect in this experiment, however, is that the subjects use the lower levels of timbre perception to guide their responses. Thus, significant overlap of classifications in the response data is expected.

The way in which the different templates are treated by the subjects is determined by the choice of template sounds. With a considerable range of timbres presented in the template for a particular test, the subject is less able (if performing the test correctly) to use an high level classificatory mode of listening, but will be forced to use a lower level timbre-space comparison as described by the range of the sounds in the template. For example, if four mid-range “standard” string tones are presented, the task that the subject is being asked to perform is different from that where the template consists of a broad range of string tones.

Secondly, the template can be used to emphasise that loudness, pitch, duration and/or environmental considerations are not of interest (where appropriate) by presenting a range of those values. This describes an “instrument space” (see Subsection 2.3.4) in the mind of the participant which the subject uses to make comparisons. By presenting a range of “non-timbral” attributes (particularly pitch and duration) it is emphasised to the subject that those properties are not of interest. This concept has been devised to counteract the problems of the non-timbral aspects dominating perception in similarity tests ([157]). Moreover, it avoids the problem of attempting to equalise the stimuli in terms of pitch, loudness, duration and environment, which is an impossible and very artificial technique

for a broad range of sounds (as discussed in Subsection 2.3.3).

As described in the text of Figure 3.1 and of Figure 3.2, the families of interest, and thus the templates, are not intended to exactly match the traditional orchestral types. Rather, they pertain to more precise characteristics. In particular, the Percussive type is not intended to represent percussion instruments, which encompass a vast range of instruments of many forms. Rather, the family is one united by a particular amplitude characteristic or playing style. Overall, the emphasis is on matching to the template, not the perceived classification associated with the title. The actual sounds of the templates are given in Table 3.1. The codes correspond to those associated with the stimuli as described in Appendix A.

3.5 Analysis of Results

After the responses of the participants were gathered as described in Section 3.4, the data was analysed using methods as described in this section. The test responses were coded into numerical values 0-3 for analysis purposes; 0 = “-like”, 1 = “-ish”, 2 = “hardly” and 3 = “not”, corresponding to increasing dissimilarity. In the following subsections the subjects themselves have been coded as follows:

1. Subject 1 is the author, and so it of interest as to whether prior knowledge of the stimuli and the experiment influenced the results in an obvious manner.
2. Subjects 1 through 7 are the “more musical” participants, comprising those participants with significant musical, or acoustical listening, experience.
3. Subjects 8 through 14 are the less- and non-musical participants.

The labels used in this section are of the form XXX-Y where XXX is the test type (STR = String, WOO = Woodwind, BRA = Brass, HAM = Hammered, PER = Percussive and SYN = Synthetic / Test-Tone) and Y is the number of the subject. A structured approach has been taken, to analyse the complex result structure from a number of related perspectives with statistical methods which reinforce each other. The mathematical techniques used are described in some more detail and in the context of similar techniques in Appendix B.

| Code | Description | Length(s) |
|-----------------------------------|-----------------------------------|-----------|
| <i>String Type</i> | | |
| 2 | violin, bowed, A4, open | 3.13 |
| 5 | violin, pizzicato A4, stopped | 0.33 |
| 25 | acoustic bass, plucked A1 stopped | 2.43 |
| 18 | cello, martele A3, stopped | 1.42 |
| <i>Woodwind Type</i> | | |
| 29 | flute, flutter A5 | 4.88 |
| 35 | oboe, A4 | 2.42 |
| 38 | E flat clarinet, A3 | 4.47 |
| 41 | bassoon, A2 | 4.80 |
| <i>Brass Type</i> | | |
| 62 | alto trombone, A4 | 2.06 |
| 66 | tuba, A2 | 4.39 |
| 67 | trombone pedal note, A1 | 2.47 |
| 52 | C trumpet, A5 | 4.28 |
| <i>Hammered Tonal Type</i> | | |
| 70 | 9' Hamburg Steinway, loud A2 | 3.07 |
| 76 | symphonic marimba A4 | 0.96 |
| 81 | glockenspiel, brass beater A5 | 1.59 |
| 72 | 9' Hamburg Steinway, loud A4 | 3.02 |
| <i>Percussive Type</i> | | |
| 84 | snare drum, hit | 0.23 |
| 101 | finger cymbals | 1.43 |
| 12 | viola, pizzicato A4, open | 0.64 |
| 107 | tambourine, pop | 0.75 |
| <i>Synthetic / Test-Tone Type</i> | | |
| 110 | synthesized analogue bass, A1 | 3.60 |
| 121 | metallic pad A1 | 1.00 |
| 135 | sine wave A4 | 1.00 |
| 142 | sawtooth A3 | 1.00 |

Table 3.1: Sounds Used as Perceptual Test Templates (in presentation order)

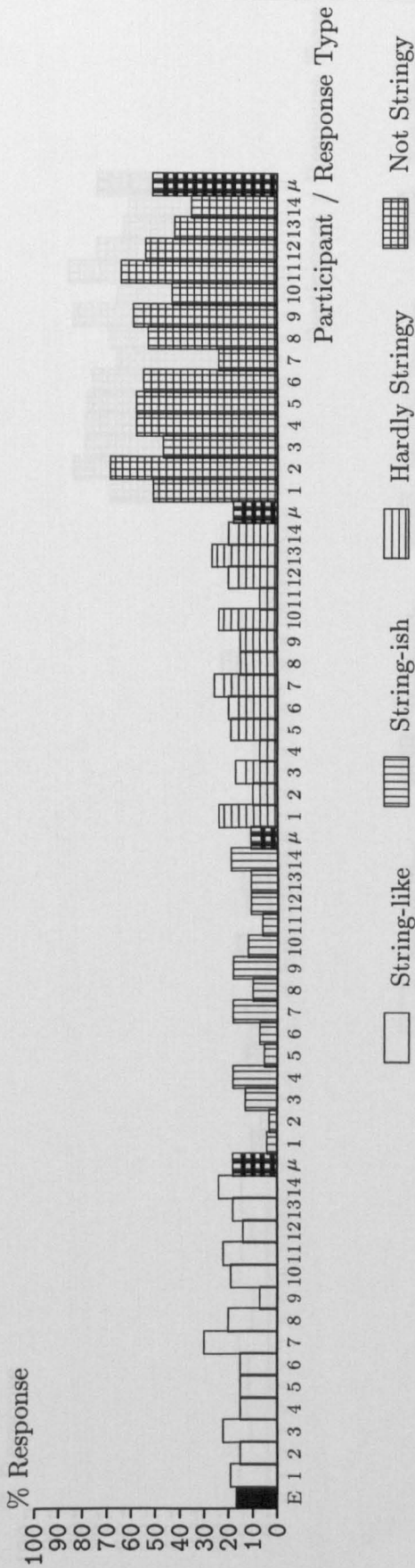


Figure 3.3: Distribution of Responses Over All Sounds for All Participants with the String Type Test

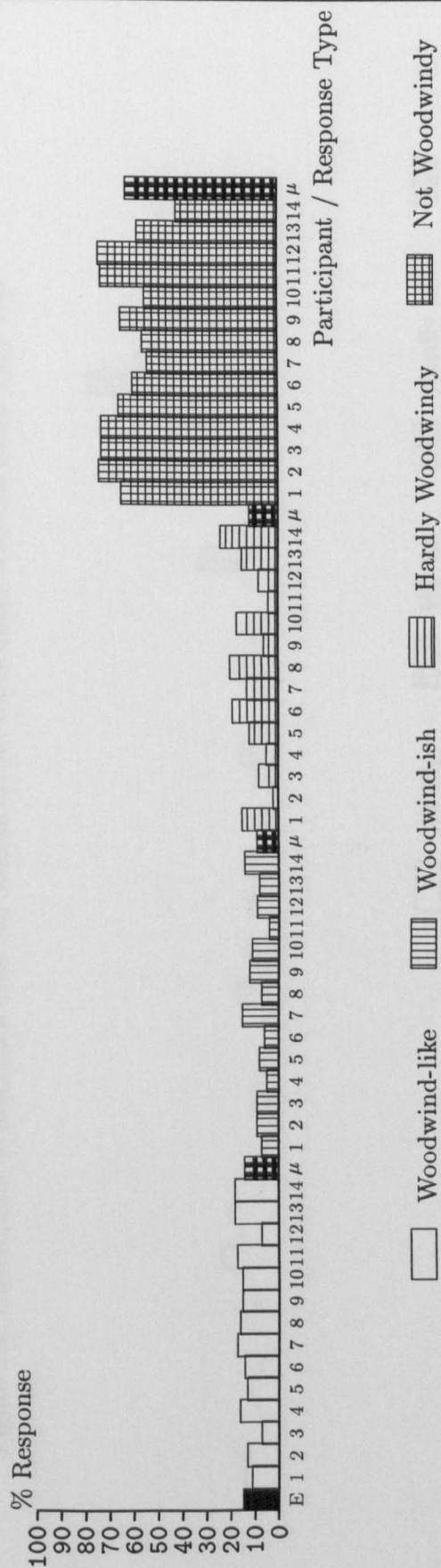


Figure 3.4: Distribution of Responses Over All Sounds for All Participants with the Woodwind Type Test

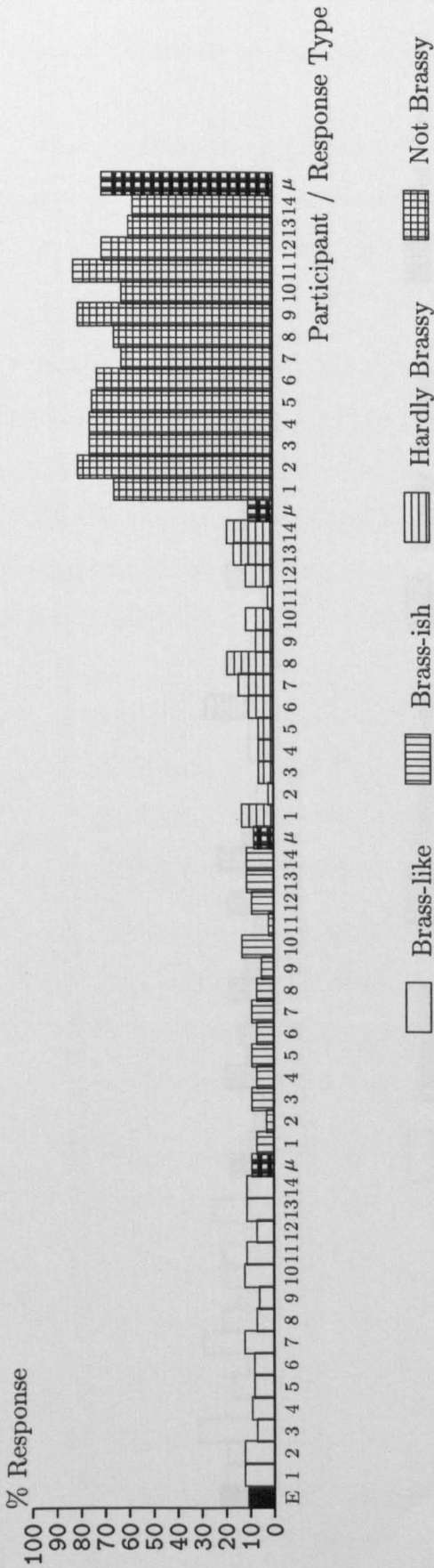


Figure 3.5: Distribution of Responses Over All Sounds for All Participants with the Brass Type Test

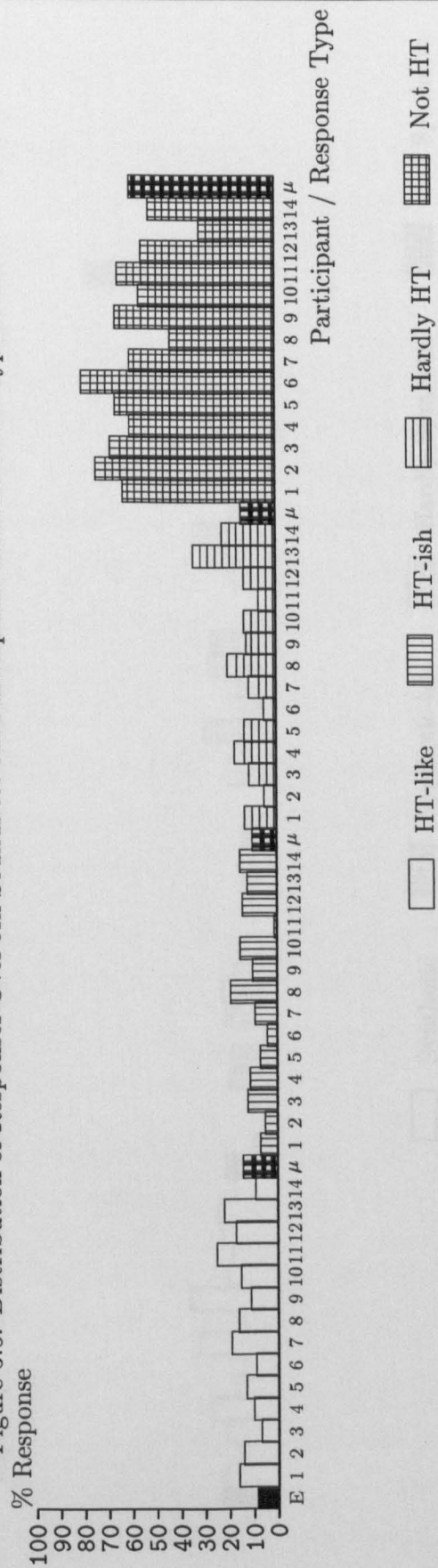


Figure 3.6: Distribution of Responses Over All Sounds for All Participants with the Hammered Tonal Type Test

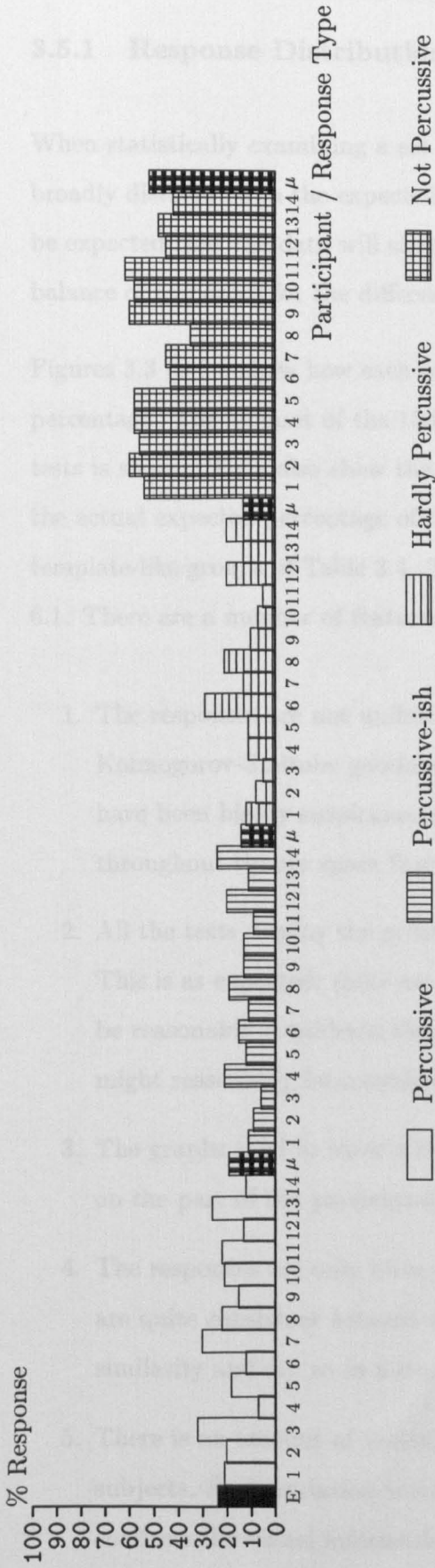


Figure 3.7: Distribution of Responses Over All Sounds for All Participants with the Percussive Type Test

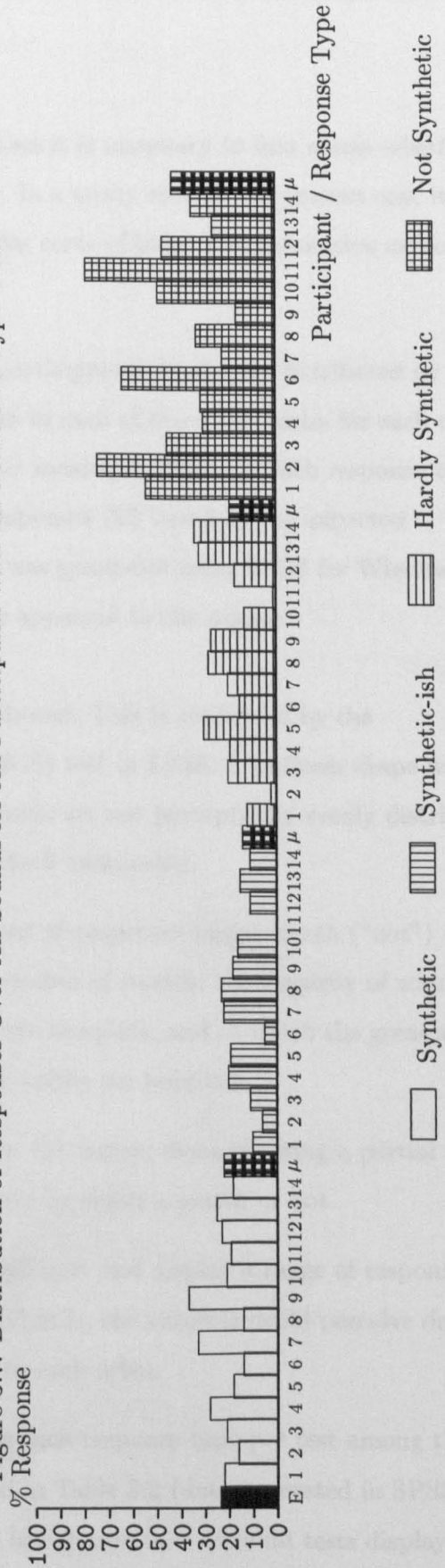


Figure 3.8: Distribution of Responses Over All Sounds for All Participants with the Synthetic/Test-Tone Type Test

3.5.1 Response Distributions

When statistically examining a set of responses it is necessary to first assess whether it is broadly distributed in the expected manner. In a study such as the present one, it might be expected that the data will show particular sorts of bias. This subsection explores the balance of responses for the different tests.

Figures 3.3 to 3.8 show how each of the 14 participants' results are distributed by percentage. The amount of the 153 responses in each of the 4 categories for each of the 6 tests is shown. They also show the arithmetic mean (μ) values for each response type, and the actual expected percentage of "-Like" responses (E) based on the expected template-like groups of Table 3.4. The data was generated using SPSS for Windows release 6.1. There are a number of features that are apparent to the author:

1. The responses are not uniformly distributed. This is confirmed by the Kolmogorov-Smirnov goodness-of-fit (K-S) test in SPSS. A uniform shape would have been highly suspicious, as the sounds are not perceptually evenly distributed throughout timbre space from any of the 6 viewpoints.
2. All the tests display the greatest number of responses in the fourth ("not") category. This is as expected; there are a large number of sounds, the majority of which cannot be reasonably considered the same as the template, and of which the greater number might reasonably be considered totally unlike the template.
3. The graphs tend to show a slight dip in the centre, demonstrating a partial tendency on the part of the participants to classify as either a match or not.
4. The responses are only binary in a small part and display a range of responses which are quite consistent *between subjects*. That is, the subjects could perceive degrees of similarity and did so in a similar way to each other.
5. There is an amount of variation within each response type per test among the subjects. Such variation is summarised in Table 3.2 (data generated in SPSS). This confirms the visual information of the bar charts; that different tests display different variation of balance of responses among the subjects. This in turn indicates the nature of the perceptual task being undertaken. For example, the Brass test generally shows the most limited variation among the subjects, indicating that they

were all judging in a similar manner; the template and task were perceptually clear. However, with more complex tests where the template and associated judgement task is more vague, the variation is stronger (such as the Synthetic test).

3.5.2 Pearson Correlation Coefficients

Correlation is a statistical method for attempting to understand the direction, form and strength of the relationship between two variables. The procedure used here is the Pearson method, which is the most common type used with interval data. Positive values indicate that variables tend to increase and decrease together, whereas negative values indicate opposite movement. The absolute magnitude of the coefficient indicates the strength of the relationship. The result ranges from -1 to +1 (perfectly opposite to perfectly similar relationship). Finally, the Pearson method finds linear correlation, and so the coefficient also indicates whether the relationship tends toward linearity.

| Statistic | Response 0 (“-like”) | Response 1 (“-ish”) | Response 2 (“hardly”) | Response 3 (“not”) |
|-----------------------------------|--------------------------------|-------------------------------|---------------------------------|------------------------------|
| <i>String Type</i> | | | | |
| Mean Average | 18.6 | 11.6 | 18.3 | 51.4 |
| Standard Deviation | 5.5 | 5.4 | 6.6 | 11.9 |
| Range | 23.5 | 15.7 | 20.3 | 45.1 |
| <i>Woodwind Type</i> | | | | |
| Mean Average | 14.5 | 9.4 | 12.3 | 63.8 |
| Standard Deviation | 3.4 | 3.3 | 6.6 | 9.8 |
| Range | 10.5 | 11.1 | 22.2 | 32.0 |
| <i>Brass Type</i> | | | | |
| Mean Average | 9.9 | 8.3 | 10.5 | 71.3 |
| Standard Deviation | 2.3 | 3.0 | 5.5 | 8.2 |
| Range | 5.9 | 10.5 | 17.0 | 25.5 |
| <i>Hammered Tonal Type</i> | | | | |
| Mean Average | 14.8 | 10.5 | 14.4 | 60.3 |
| Standard Deviation | 5.2 | 5.0 | 7.4 | 12.3 |
| Range | 18.3 | 18.3 | 28.8 | 49.0 |
| <i>Percussive Type</i> | | | | |
| Mean Average | 19.8 | 14.4 | 13.6 | 52.2 |
| Standard Deviation | 7.0 | 5.2 | 7.0 | 8.5 |
| Range | 24.2 | 18.3 | 24.2 | 26.8 |
| <i>Synthetic / Test-Tone Type</i> | | | | |
| Mean Average | 22.5 | 14.0 | 19.7 | 43.8 |
| Standard Deviation | 6.9 | 6.9 | 10.3 | 19.3 |
| Range | 23.5 | 23.5 | 32.7 | 63.4 |

Table 3.2: Basic Statistics of Percentage Responses for Perceptual Tests Among the Participants

Table with 27 columns (SYN-1 to SYN-27) and 100 rows (WOO-1 to SYN-14). The table contains Pearson Correlation Coefficients for various perceptual study test results. The values range from approximately -0.410 to 0.819.

CONTINUED →

Table 3.3(cont.): Pearson Correlation Coefficients of Perceptual Study Test Results

Table 3.3 presents the coefficients for all 14 subjects and all 6 tests. The values were generated using Microsoft Excel version 5.0. The correlation table has been reproduced in full to allow the reader to examine the detail. Attempting to summarise the data numerically would compromise some of the nuances that are present. The author interprets the results as follows:

1. The values corresponding to the same test with different participants are generally high or very high (where high is >0.65) indicating that the subjects were acting in a similar manner as they performed the tests. This implies the following:
 - (a) All types of participant, whatever their background, were capable of understanding the concepts involved and performing the matching task. That is, all types of participant had the ability to perceive timbral similarity, and did so in a similar way.
 - (b) The template matching method provides a stable mental percept which can be used in judgement. Otherwise, the responses would not be so highly correlated.
2. Coefficients of *different* tests display similar correlations with different pairs of subjects. That is, not only is there correspondence between different subjects performing the same test, but also coherence between tests of different timbre types:
 - (a) Different participants are using similar processes of timbre judgement.
 - (b) There are systematic relations between timbre types in perception.
3. The results for the Hammered and Percussive types often demonstrate significant positive correlation between them, for the same subject and between participants. This reflects the similarity of their definitions, particularly with regard to the attack phase of their templates. This also shows that although the templates were not composed of the same sounds, their perceived timbral similarity was apparent:
 - (a) The subjects were able to perceive the similarity of the timbral subspaces described by the templates.
 - (b) Again, this demonstrates systematic and structured perception.
4. The Woodwind and Brass types often demonstrate mild, and sometimes more major, positive correlation. This follows the known timbral similarities to an extent, which results from the physical similarities of the sources.

5. The Synthetic / Test-tone tests tend toward lower correlation values with the same test between subjects, reflecting the less precise nature of the test and its wider timbral subspace of the template.
6. The Percussive and Hammered types tend to produce significant negative correlation with the Woodwind and Brass types. This indicates the nature of the timbre types involved; in particular the effect of the attack type.
7. The Synthetic test results quite often generate negative correlations with other types of moderate magnitude. Yet the interpretation of that which is synthetic seems to be different for different subjects. For example, SYN-5 has notable negative correlation with Percussive tests; whereas subjects 3,4,7, and 9 have similar opposition to Strings.

As might be expected, there is some variability from the general pattern as described above. However, it is hard to find any systematic differences of statistical significance from the correlation analysis between subjects relating to musical background, or prior knowledge, indicating the inherent human ability to comprehend the structure of timbral difference.

3.5.3 Modal Responses for Individual Cases

This subsection considers how the responses relate to individual cases based on modal averages. These indicate the most frequently chosen responses for each stimulus by the 14 participants, for each of the 6 tests. Such results indicate both how well the template matching scheme worked and also whether responses to stimuli are as expected or not. Based on the results from Subsection 3.5.2, it seems reasonable to the author to consider the mode values taken from *all* subjects, not just those of a particular musical background. This is because all the participants were able to judge perceptual similarity in a similar way, and so all responses are valid in finding an average.

| <i>String Type</i> | |
|----------------------------|---|
| Actual Template | 2, 5, 18, 25 |
| Template-Like Group | 1-27 |
| Modal Resp. 0 (“-like”) | 1-16, 18-26, 115, 116, 125 |
| Modal Resp. 1 (“-ish”) | 17, 27, 36, 39, 40, 68-70, 80, 112, 121, 124 |
| Modal Resp. 2 (“hardly”) | 35, 41-44, 52, 64, 66, 71-73, 110, 111, 119, 123, 126, 128, 130 |
| Modal Resp. 3 (“not”) | 28-34, 37, 38, 45-51, 53-63, 65, 67, 74-79, 81-109, 113, 114, 117, 118, 120, 122, 127, 129, 131-153 |
| <i>Woodwind Type</i> | |
| Actual Template | 29, 35, 38, 41 |
| Template-Like Group | 28-51 |
| Modal Resp. 0 (“-like”) | 28-45, 47, 48, 51, 58, 63, 64, 120, 126 |
| Modal Resp. 1 (“-ish”) | 46, 49, 56, 57, 65-67, 115 |
| Modal Resp. 2 (“hardly”) | 21, 60-62, 119, 151 |
| Modal Resp. 3 (“not”) | 1-20, 22-27, 50, 52-55, 59, 68-114, 116-118, 121-125, 127-150, 152, 153 |
| <i>Brass Type</i> | |
| Actual Template | 52, 62, 66, 67 |
| Template-Like Group | 52-67, 151 |
| Modal Resp. 0 (“-like”) | 43, 52-63, 65-67, 151 |
| Modal Resp. 1 (“-ish”) | 35, 36, 40-42, 46, 48-50, 120, 126 |
| Modal Resp. 2 (“hardly”) | 21, 38, 39, 44, 45, 47, 51, 64, 119 |
| Modal Resp. 3 (“not”) | 1-20, 22-34, 37, 68-118, 121-125, 127-150, 152, 153 |
| <i>Hammered Tonal Type</i> | |
| Actual Template | 70, 72, 76, 81 |
| Template-Like Group | 68-79, 81-83 |
| Modal Resp. 0 (“-like”) | 17, 27, 68-79, 81-83, 93, 101, 103, 113, 117, 122, 125, 129, 130 |

CONTINUED →

Table 3.4: Stimulus Codes Corresponding to Particular Modal Responses for all Participants

| | |
|-----------------------------------|---|
| Modal Resp. 1 (“-ish”) | 26, 89, 98-100, 116, 118, 123, 124, 127 |
| Modal Resp. 2 (“hardly”) | 5, 23, 25, 85-88, 90, 91, 94-97, 104, 107, 109, 112 |
| Modal Resp. 3 (“not”) | 1-4, 6-16, 18-22, 24, 28-67, 80, 84, 92, 102, 105, 106, 108, 110, 111, 114, 115, 119-121, 126, 128, 131-153 |
| <i>Percussive Type</i> | |
| Actual Template | 12, 84, 101, 107 |
| Template-Like Group | 5, 6, 12, 17, 23, 76-104, 107, 109, 152 |
| Modal Resp. 0 (“-like”) | 5-7, 12, 74, 77, 78, 81, 82, 84-94, 98-101, 103, 104, 107, 109, 116, 117, 127, 130, 152 |
| Modal Resp. 1 (“-ish”) | 8, 17, 24-27, 71, 73, 75, 76, 79, 83, 95-97, 102, 113, 118, 122, 124, 125, 128, 129 |
| Modal Resp. 2 (“hardly”) | 13, 23, 50, 68, 69, 72, 80, 112, 121, 123 |
| Modal Resp. 3 (“not”) | 1-4, 9-11, 14-16, 18-22, 28-49, 51-67, 70, 105, 106, 108, 110, 111, 114, 115, 119, 120, 126, 131-151, 153 |
| <i>Synthetic / Test-Tone Type</i> | |
| Actual Template | 110, 121, 135, 142 |
| Template-Like Group | 110-146 |
| Modal Resp. 0 (“-like”) | 76, 108-111, 117-119, 121, 122, 124, 127-129, 131-147, 149, 150, 152 |
| Modal Resp. 1 (“-ish”) | 27, 41, 42, 57, 64, 66, 77-80, 113, 123, 125, 126 |
| Modal Resp. 2 (“hardly”) | 23, 36, 43, 44, 51, 55, 56, 60, 62, 65, 81, 89, 112, 130 |
| Modal Resp. 3 (“not”) | 1-22, 24-26, 28-35, 37-40, 45-50, 52-54, 58, 59, 61, 63, 67-75, 82-88, 90-107, 114-116, 120, 148, 151, 153 |

Table 3.4: Stimulus Codes Corresponding to Particular Modal Responses for all Participants

Table 3.4 presents the modal averages for all 14 subjects and all 6 tests. The values were generated using Microsoft Excel. The stimulus codes correspond to the stimuli as listed in Appendix A. The template-like groups listed in the table are those stimuli which were supposed, at the beginning of the experiment, to be represented by the actual template sounds. That does not mean that it was expected that they were the *only* sounds which could result in a response average equal to 0, but that they form a core set of stimuli which were regarded initially to have considerable similarity to the actual template sounds. The

author interprets the results as follows:

1. The actual template stimuli have all been matched correctly, on average, as they all appear in the list of modal response 0 for their respective tests. Although as expected, this shows that the participants were capable of maintaining the mental images of the timbral templates during the experiment and applying the memory correctly later, as the subjects rarely asked for the template sounds to be repeated more than a couple of times per test.
2. A more important result is that the template-like groups are generally matched very well. This shows that the mental image (timbre space) describing the range of timbral form, as imparted from the sonic information in the actual template group, has been successfully used to match the features of stimuli which fit into that mental timbre space. This shows that the template technique achieves its aim of allowing similarity comparisons with timbral spaces, rather than one-to-one matches only, and thus matching a wide range of features and avoiding the effects of pitch, loudness and duration on the timbre comparisons.
3. For the String type, it is interesting that sound 17 does not have a mode response of 0, which shows how deceptive a single instance of sound can be. The position of 115, 116 and 125 in the 0 set seems logical, despite their synthesised nature. The inclusion of the woodwinds 36, 39, 40, and 124 in the modal 1 set seems unusual, but the low piano notes (68-70) less so. The bowed vibraphone (80) is a particularly good example of how some of the elements of the sound have been matched nicely with the template timbre space by the participants.
4. For the Woodwind type, stimulus 50 (alto saxophone scream) has dropped into the “not” category, which is particularly noteworthy; it also has a mode of 1 for the Brass type. That 46 and 49 are in the mode 1 group for both the Woodwind and Brass types emphasises how saxophones may be perceived as somewhat between the two types. Additional confusion between the Woodwind and Brass is emphasised by 58, 63, and 64 in mode 0 and 56, 57 and 65-67 in the mode 1 set. That 120 and 126 are considered Woodwind-like is logical, though.
5. With the Brass type, only 64 (muted tenor trombone) of the template-like group is considered rather less Brass-like than might be expected. 43 (bass saxophone) is

- present in the mode 0 group as well. That which is true for the Brass in the Woodwind test is true in reverse for the Brass test, with a considerable number of Woodwinds in the mode 1 set.
6. With the Hammered Tonal test, stimulus 80 (bowed vibraphone, mentioned previously in connection with the String type) is not recognised as a Hammered Tonal type instrument being played in an unusual manner. This seems logical, as the bowing action is a strong cue and prevents a reasonable match with the template description. A considerable number of other, perceptually accurate matches appear in the mode 0 set which have a struck, yet tonal character which were not anticipated in the original template-like group. There is a visible progression of character through to the mode 3 set, as is expected, assuming the participants performed the test as requested.
 7. The Percussive type is different from the other tests, being concerned particularly with the shape of the amplitude curve, rather than the overall complex. This also has an added dimension in that the group is not *percussion*, which has a very broad definition, but is more precise. It would seem that the participants performed the test as required, however, despite the additional complexities. Those sounds which are not necessarily percussion instruments, but have the Percussive type nature have low modal values (hard string sounds, higher piano notes, many of the percussion instruments), but not the percussion instruments which do not fit the template (80, 105, 106, 108).
 8. The Synthetic type was the test with the least predictable outcome before the testing, due to the complexity of the concept. As expected, the “clearly” synthetic test tones (131-146) were successfully matched by the subjects, as were quite a few of the wavetable synthesised sounds (111-130). The participants were unconvinced by some of the miscellaneous sounds’ “natural” qualities (147, 149, 150, 152) and quite a few of the actual instrument samples (in the mode 1 set). Overall, it is very difficult to say what makes a sound appear synthetic from this test alone, with several unquantified effects present, such as recording quality.

It is consistently apparent from the above points that the auditory cortex is quite capable of matching qualities of sounds to mental timbre spaces which have been constructed from

the template sounds. It may also be that humans normally identify sounds based on previous experience of a region of timbre space, using a combination of sound instances to build up an identification picture. This would explain some of the clear confusions in the response data which would not be likely were there more information available than single sound instances:

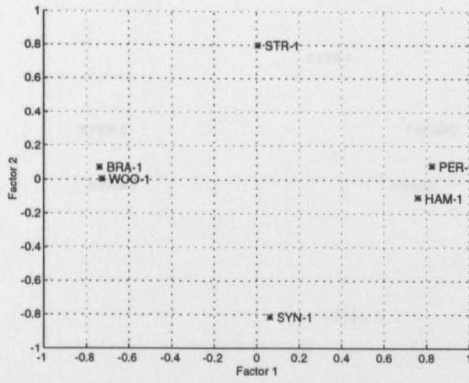
“... the human will do the best he can with what is presented to him, within the limits defined by the input and his modes of categorizing.” [51]

What is particularly interesting is that the modal responses show that the participants are consistent with each other in their assessments. That is, the confusions that are apparent are a reasonable assessment of the information; on average 14 subjects have perceived the data in a similar way. It is also worth noting that the modal responses are not chance values resulting from an highly distributed set of responses; the results above are backed up by other average metrics.

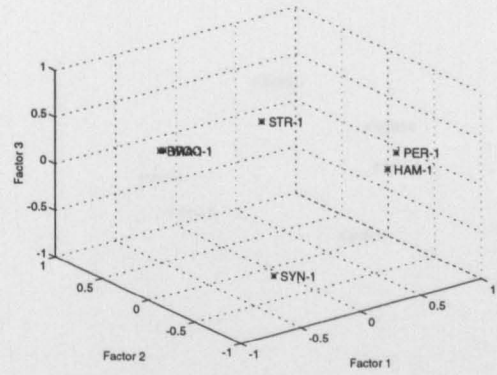
The results also show that there is a consistent structure in timbre perception. That is, the relationships between the groups to which the stimuli may be said to belong (strings, woodwind, brass and so on) are logically displayed in the perceptual similarity rating averages of the individual stimuli. These also demonstrate a continuous range of timbre perception, rather than abrupt classifications, and overlaps between groups of stimuli in the mental timbre space.

3.5.4 Analysis of Test Relationships Through Scaling

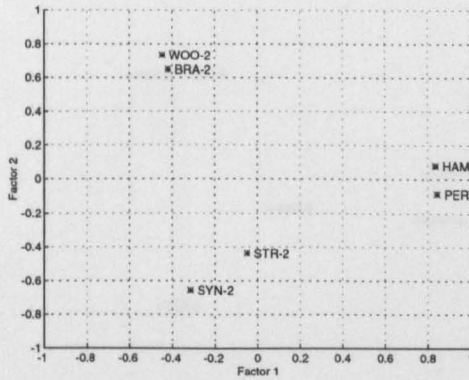
Statistical methods exist to attempt to find factors which represent the strongest contributors to variation among a group of variables. By their nature, these scaling methods can reduce the number of variables required to convey the majority of the information. This is achieved by compromising absolute accuracy. Such methods are particularly useful for visualisation purposes. SPSS has a number of these procedures, with several variations in method.



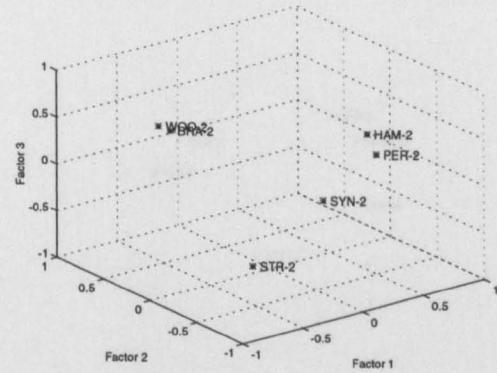
(1) Subject 1 : 2D (61.0% of Variance)



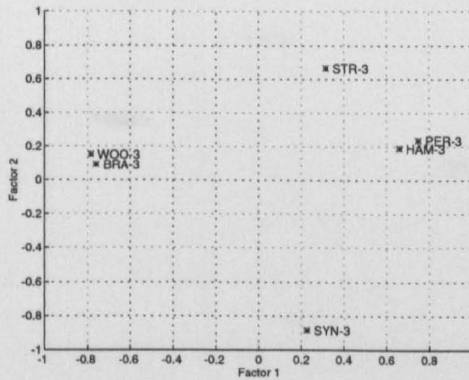
(2) Subject 1 : 3D (78.0% of Variance)



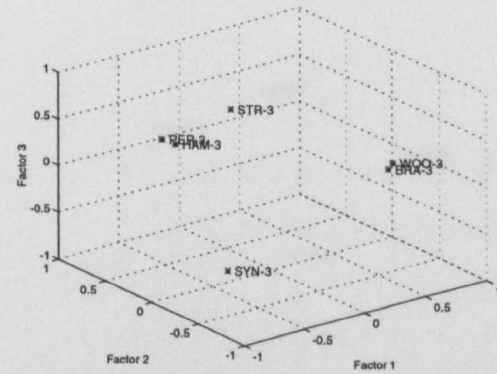
(3) Subject 2 : 2D (58.0% of Variance)



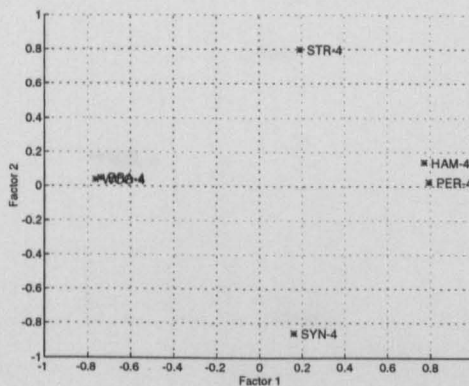
(4) Subject 2 : 3D (77.3% of Variance)



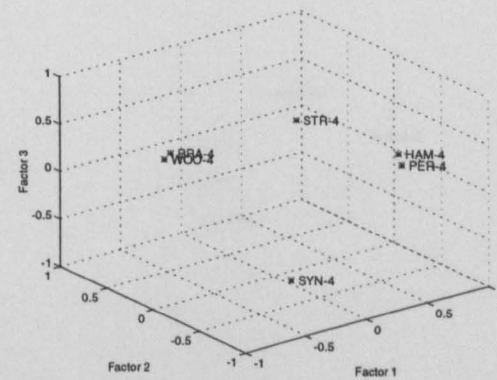
(5) Subject 3 : 2D (61.5% of Variance)



(6) Subject 3 : 3D (77.1% of Variance)

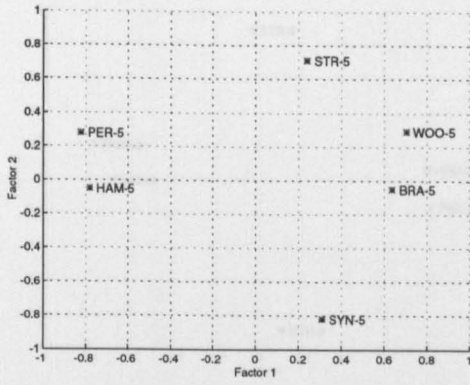


(7) Subject 4 : 2D (63.6% of Variance)

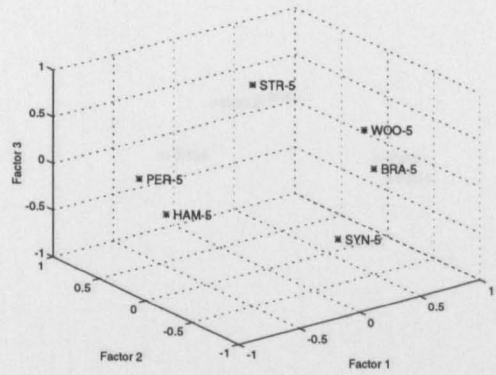


(8) Subject 4 : 3D (79.4% of Variance)

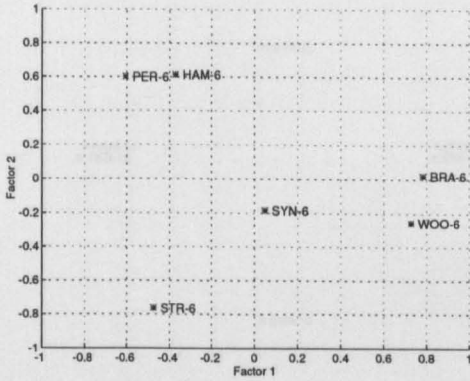
Figure 3.9: Varimax-Rotated Variable Loadings for PCA of Subjects' Test Results



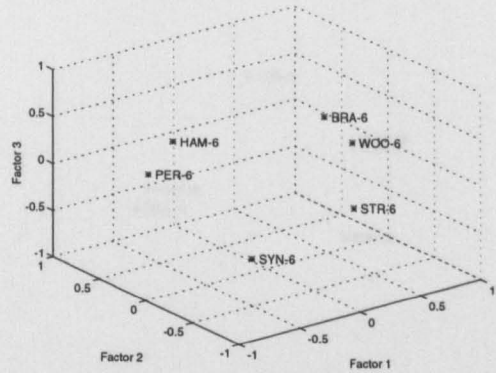
(9) Subject 5 : 2D (61.5% of Variance)



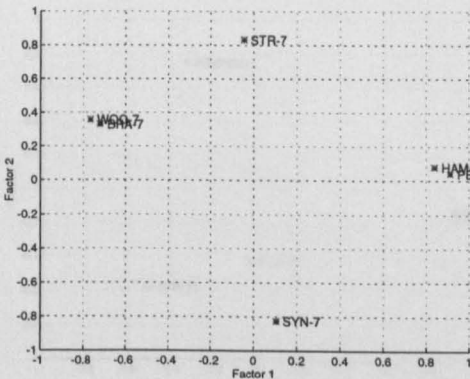
(10) Subject 5 : 3D (77.9% of Variance)



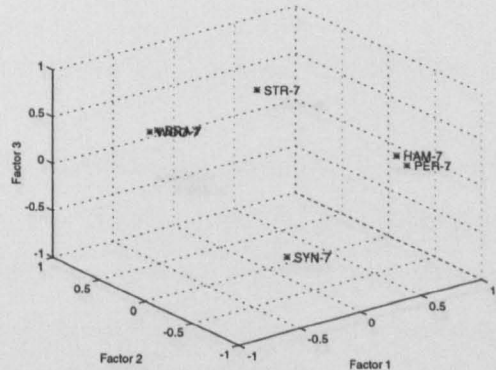
(11) Subject 6 : 2D (54.9% of Variance)



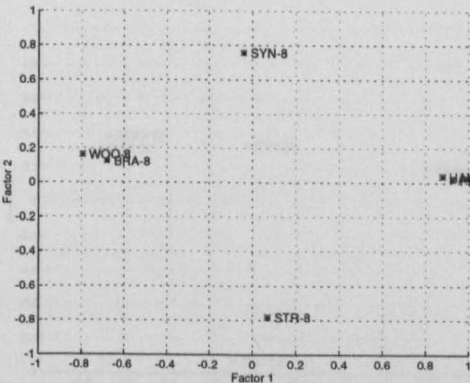
(12) Subject 6 : 3D (73.9% of Variance)



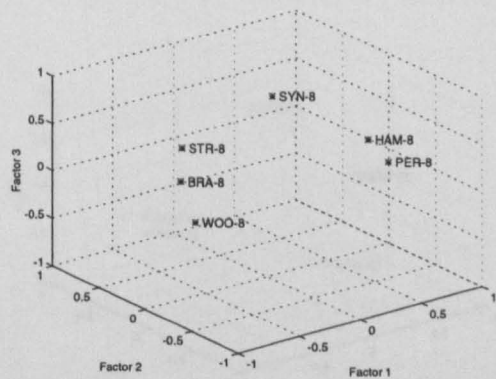
(13) Subject 7 : 2D (71.2% of Variance)



(14) Subject 7 : 3D (81.4% of Variance)

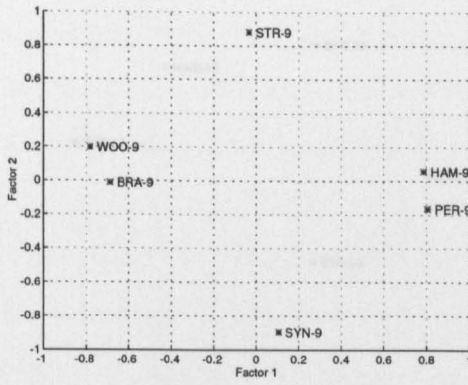


(15) Subject 8 : 2D (65.9% of Variance)

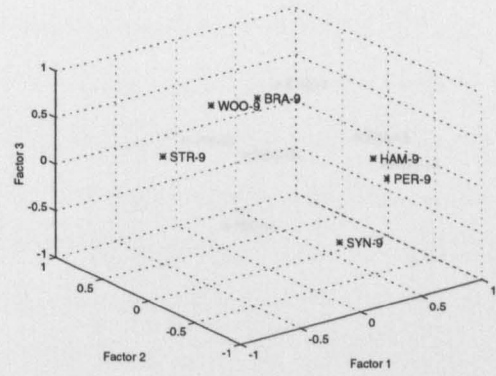


(16) Subject 8 : 3D (80.2% of Variance)

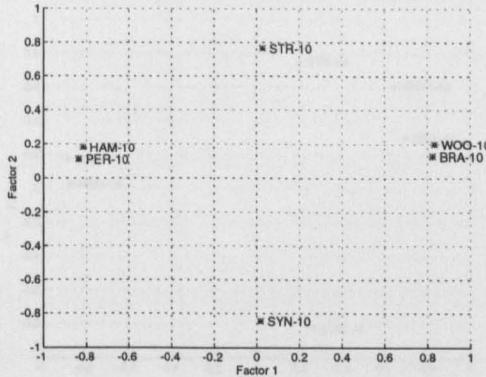
Figure 3.9(cont.): Varimax-Rotated Variable Loadings for PCA of Subjects' Test Results



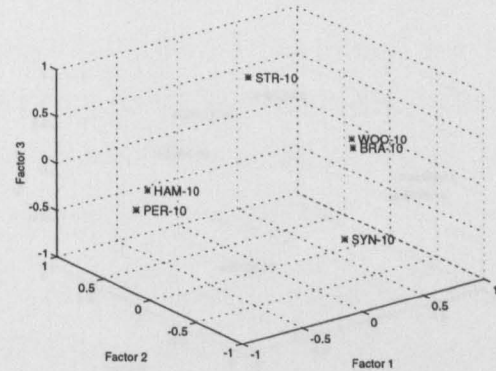
(17) Subject 9 : 2D (66.9% of Variance)



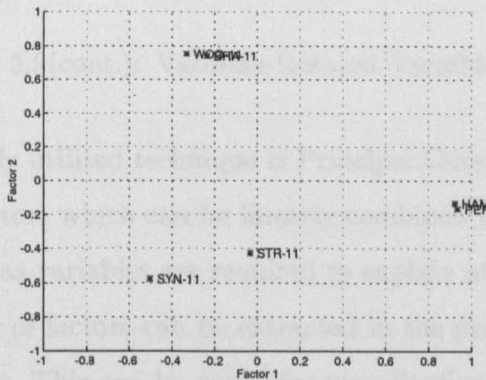
(18) Subject 9 : 3D (83.5% of Variance)



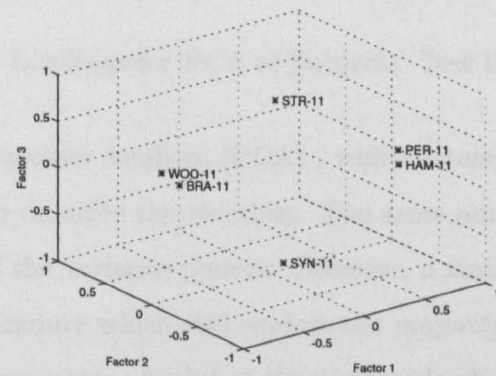
(19) Subject 10 : 2D (69.0% of Variance)



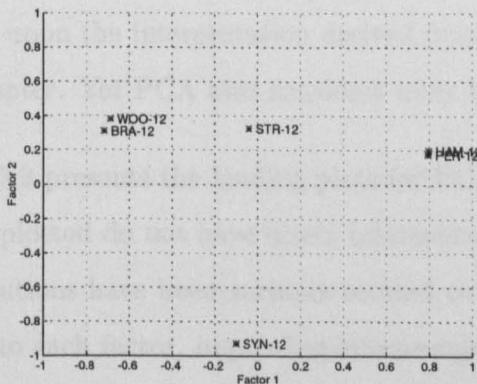
(20) Subject 10 : 3D (82.9% of Variance)



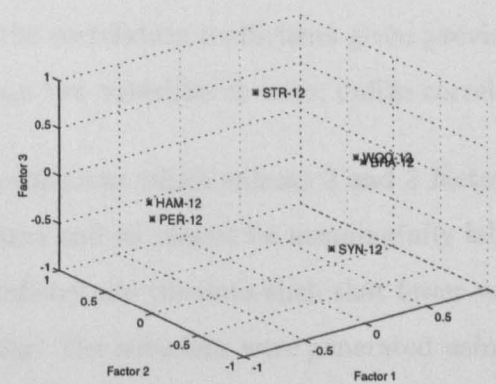
(21) Subject 11 : 2D (63.8% of Variance)



(22) Subject 11 : 3D (82.4% of Variance)

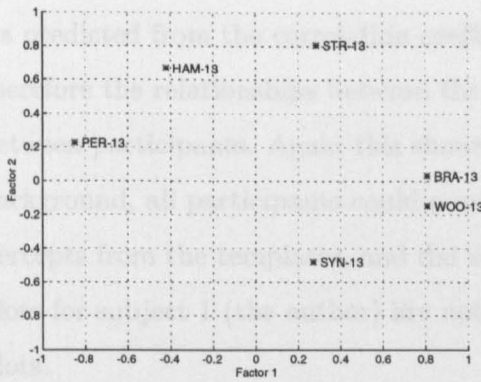


(23) Subject 12 : 2D (59.0% of Variance)

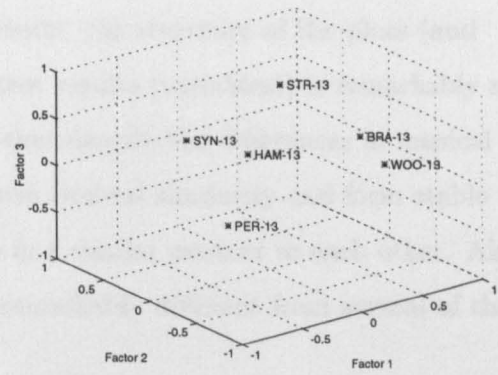


(24) Subject 12 : 3D (77.2% of Variance)

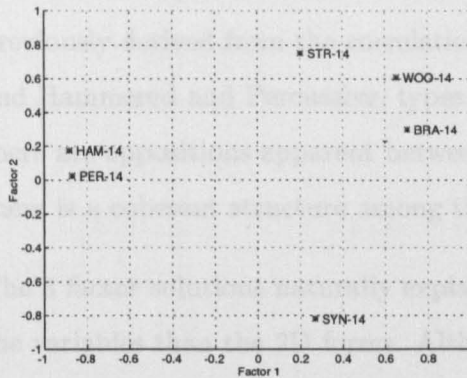
Figure 3.9(cont.): Varimax-Rotated Variable Loadings for PCA of Subjects' Test Results



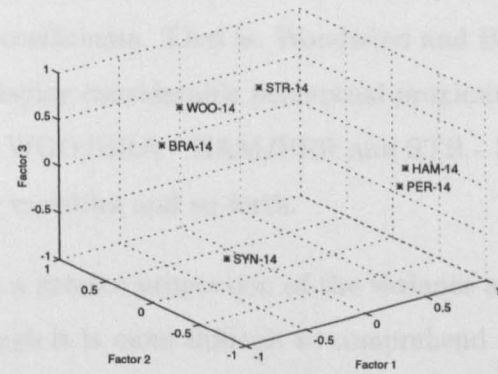
(25) Subject 13 : 2D (62.7% of Variance)



(26) Subject 13 : 3D (79.1% of Variance)



(27) Subject 14 : 2D (71.5% of Variance)



(28) Subject 14 : 3D (83.0% of Variance)

Figure 3.9(cont.): Varimax-Rotated Variable Loadings for PCA of Subjects' Test Results

A widely utilised technique is Principal Components Analysis (PCA), which attempts to find factors which can be linearly combined to describe the variables. The same number of factors as variables are required to explain *all* the variance present. However, a smaller number of factors can be extracted in the procedure which still explain the *majority* of variance. This can be useful for visualisation purposes. A plot of the variable loadings describes the relationship between the variables in this context, and thus can confirm and expand upon the interpretation derived from the correlation coefficients given previously in this chapter. Yet PCA also considers more than two variables at once, unlike correlation.

Figure 3.9 presents the loading plots for PCA solutions which extract 2 and 3 factors. The factors plotted do not have direct interpretations and so cannot be meaningfully labelled. The solutions have been varimax rotated to redistribute the data such that fewer variables load onto each factor, improving interpretability. The solutions were generated using SPSS and plotted under MATLAB version 4.2. Note that the axes do not correspond exactly between subjects. The author interprets the results as follows:

1. As predicted from the correlation coefficients, the structure of the plots (and therefore the relationships between the test results (variables)) is remarkably similar between participants. Again this shows that despite the differences in musical background, all participants could perceive timbral similarity and form stable mental percepts from the templates, and did so in a similar manner to each other. Also, the plots for subject 1 (the author) are not remarkably different from several of the other plots.
2. The relationship between the variables has large similarities with the information previously derived from the correlation coefficients. That is, Woodwind and Brass, and Hammered and Percussive, types display considerable perceptual proximity; there are oppositions apparent between WOO/BRA - HAM/PER and STR - SYN; there is a coherent structure among the variables and so forth.
3. The 3 factor solutions naturally explain a greater proportion of the variance among the variables than the 2D forms. Although it is more difficult to comprehend a 3D graph on paper than a 2D one, it is apparent that there are subtle differences between the two forms due to the difference in explained variation. The average difference in explained variance between the 2D and 3D versions is 15.9%, which is a little under a sixth of the total variance. Some pairs display the additional variation as a distinguishing effect between the similar pairs of WOO/BRA and HAM/PER (for example, subjects 8 and 10). Sometimes the additional variation appears to relate to the difference between SYN and STR (subjects 2 and 11, for example). In general, though, the difference is spread around the variables.
4. It is interesting that explained variance for the solutions is slightly higher for the less musical subjects: on average 61.7% and 77.9% for 2D and 3D for subjects 1-7; 65.5% and 81.2% for subjects 8-14. This hints that the musical subjects may have picked up on more subtle nuances of the sounds, which have been incorporated into the results, making a lower dimensional space less effective in characterising their responses.

3.5.5 Analysis of Perceived Stimulus Relationships Through Scaling

In Subsection 3.5.4, the relationship between the tests scaled through PCA to 2 and 3 dimensions is explored. This subsection considers how the factor scores for the stimuli (that is, the cases in the PCA as opposed to the variables) are arranged in the rotated factor space. In Subsection 3.5.4 the general similarities of the perceived structure between participants is demonstrated, through the variable loadings. However, PCA also allows the examination of the arrangement of the stimuli which results from those loadings. This again uses *all* the tests at once, which facilitates the consideration of the composite arrangement. The technique used in this part is the same as for the variables; that is, varimax-rotated PCA with 3 extracted factors. SPSS was used for the analysis and the results were plotted in MATLAB.


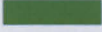


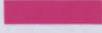

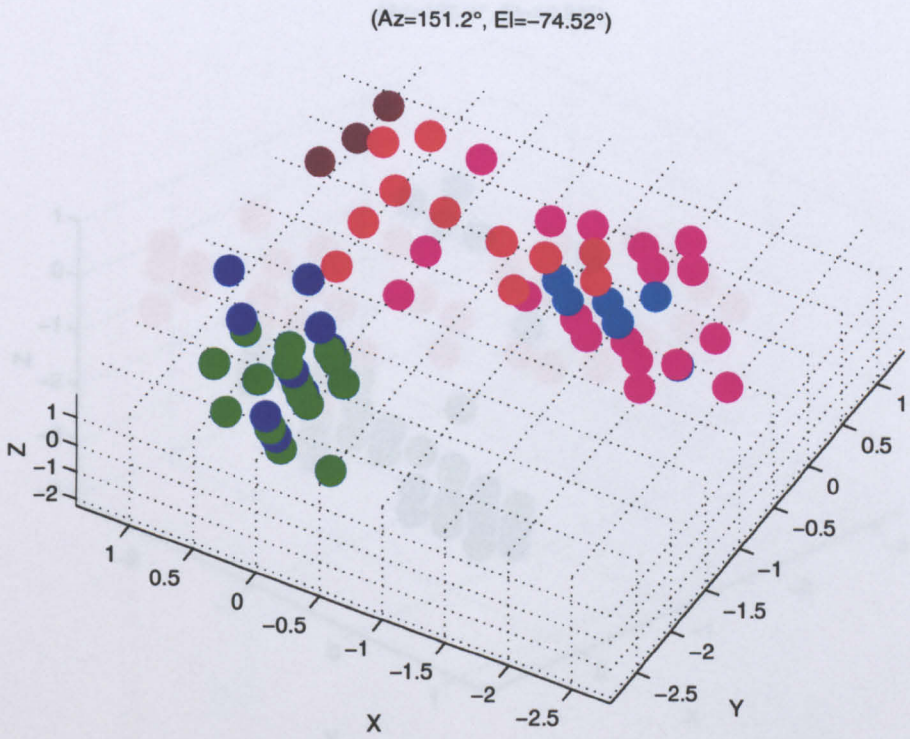
| Stimulus Code Range | Group Name | Colour Code |
|---------------------|------------|---|
| 1 - 27 | STRINGS |  |
| 28 - 51 | WOODWIND |  |
| 52 - 67, 151 | BRASS |  |
| 68 - 75 | PIANO |  |
| 76 - 109 | PERCUSSION |  |
| 131 - 145 | TEST TONES |  |

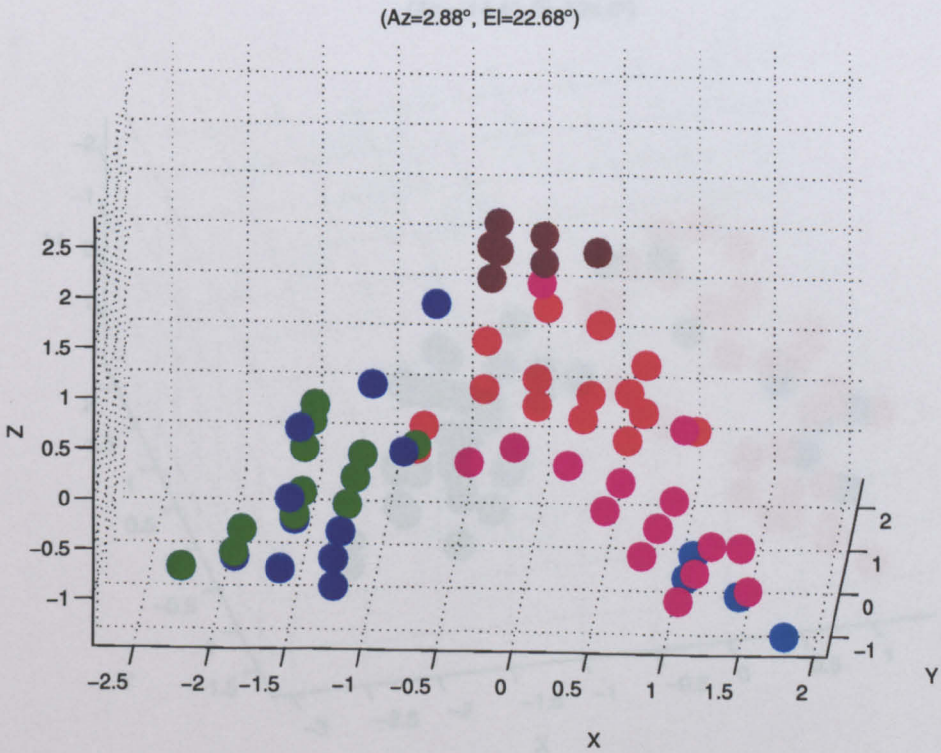
Table 3.5: Stimulus Groupings and Colour Coding used in Subsection 3.5.5

This subsection considers a set of 4 subjects; subject 1, the author; subject 6, who is a specialist in music and aural technology; and subjects 9 and 12, who are typical non-musicians. This set of subjects shows the typical similarities and variation among the full set of 14. For present purposes, groups of stimuli are formed and colour coded as described in Table 3.5, which represent the basic orchestral groups with the addition of PIANO and simple TEST TONES. A different typeface has been used for the group names to distinguish them from the more strictly defined titles used for the perceptual tests; those refer to the coherent timbre spaces described by the templates, these are “traditional” groupings of sounds. In particular, Percussive is not PERCUSSION; PERCUSSION encompasses a vast range of instrumental forms, whereas Percussive is limited by amplitude envelope, as described in the instructions on the response form (Figure 3.2). It is also notable that the groups do not include the more complex and obscure stimuli, for clarity.

Figure 3.10: Large Displacement of Stimuli in the Rotated Factor Space (PCA of Subjects' Responses)

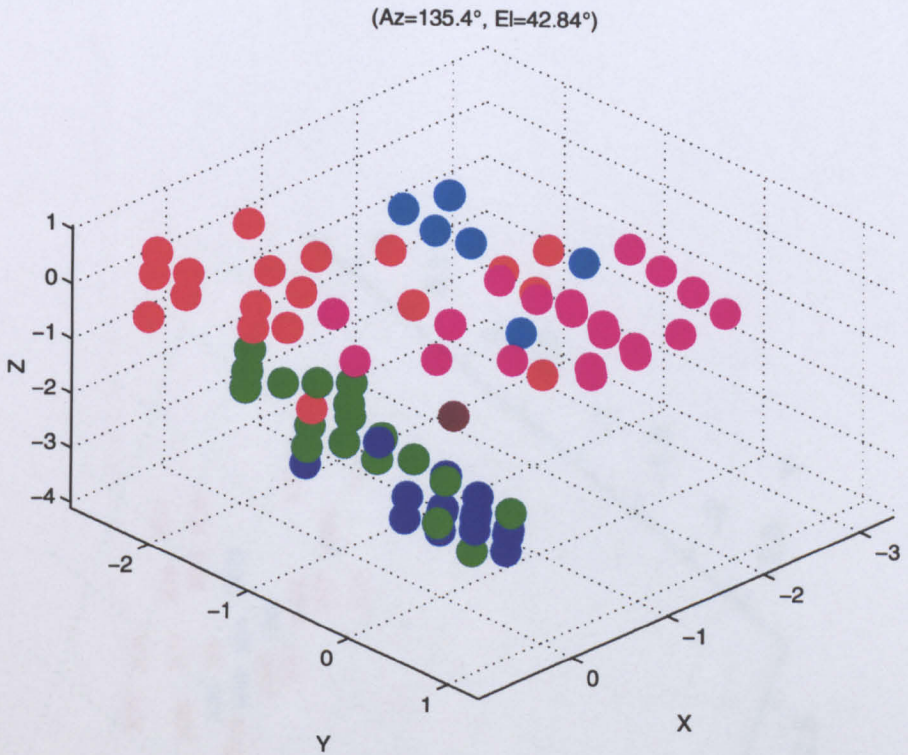


(1) Subject 1 (Author / Musician)

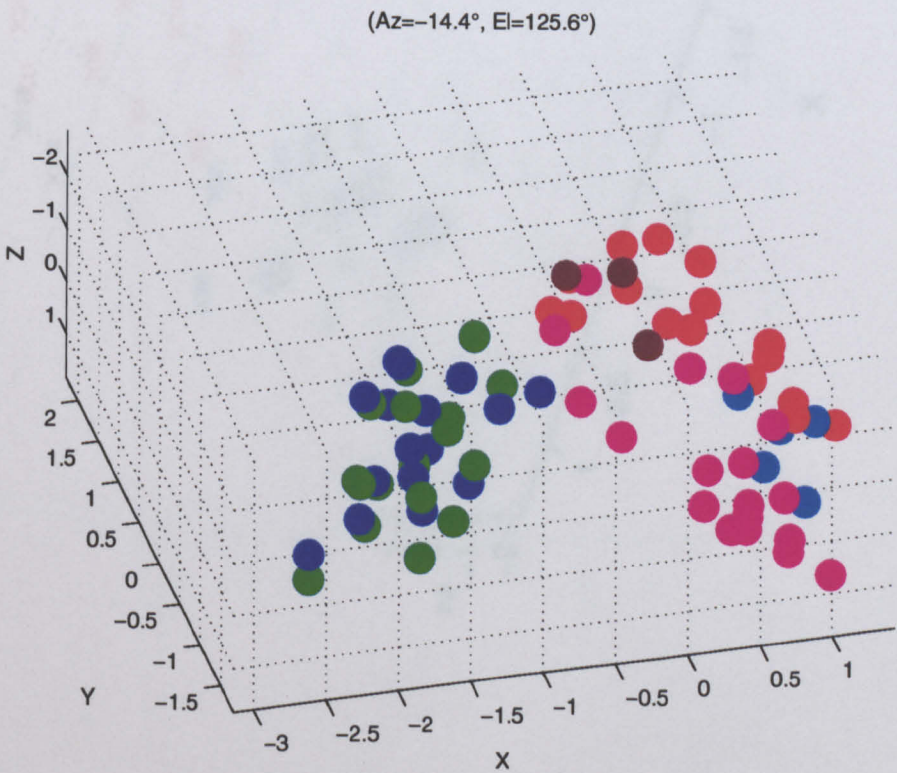


(2) Subject 6 (Musician)

Figure 3.10: Large Dot Plots of Factor Scores for Individual Stimuli Resulting from PCA of Subjects' Responses



(3) Subject 9 (Non-Musician)



(4) Subject 12 (Non-Musician)

Figure 3.10(cont.): Large Dot Plots of Factor Scores for Individual Stimuli Resulting from PCA of Subjects' Responses

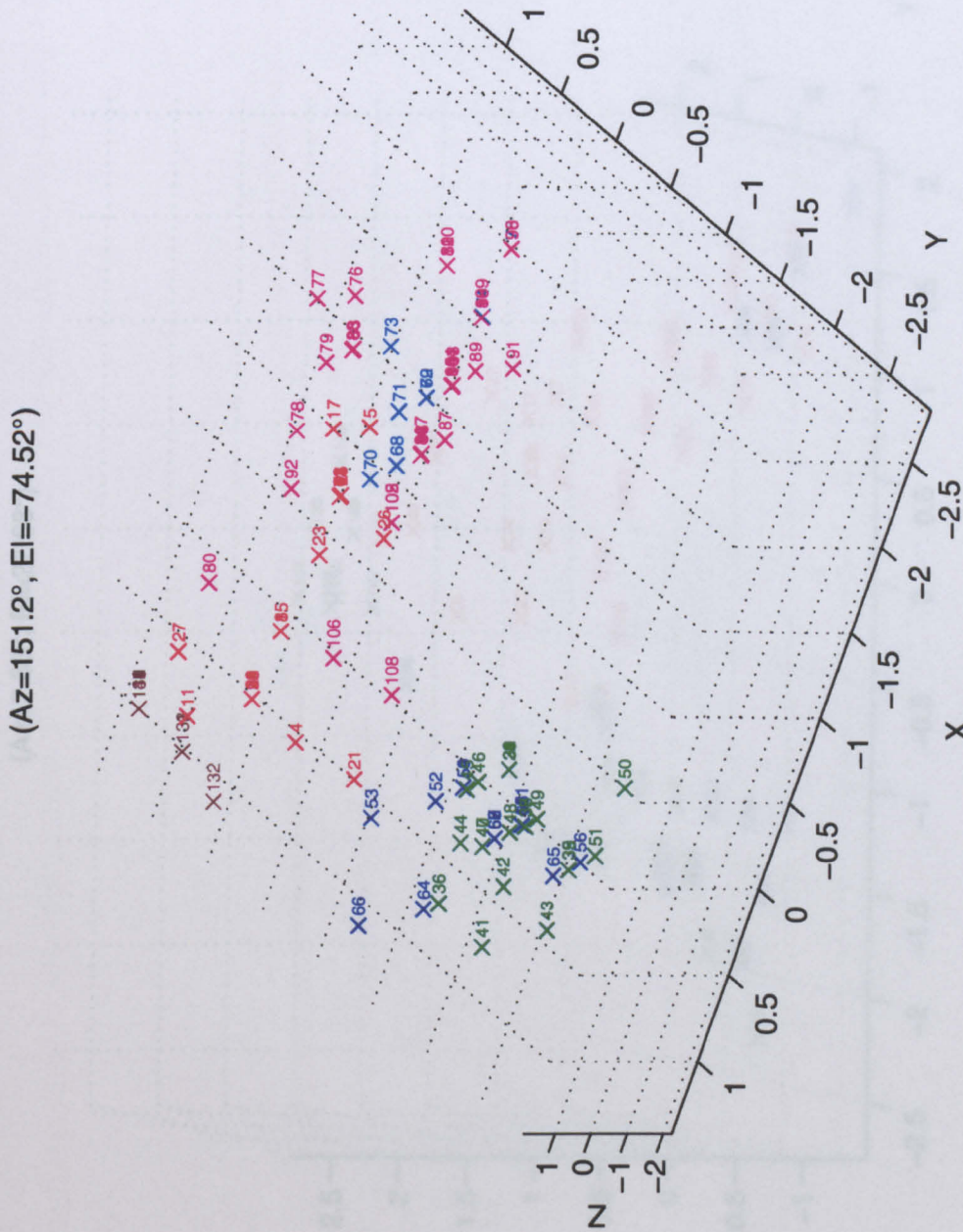


Figure 3.11: Labelled Plot of Factor Scores for Individual Stimuli Resulting from PCA of the Responses for Subject 1 (Musician)

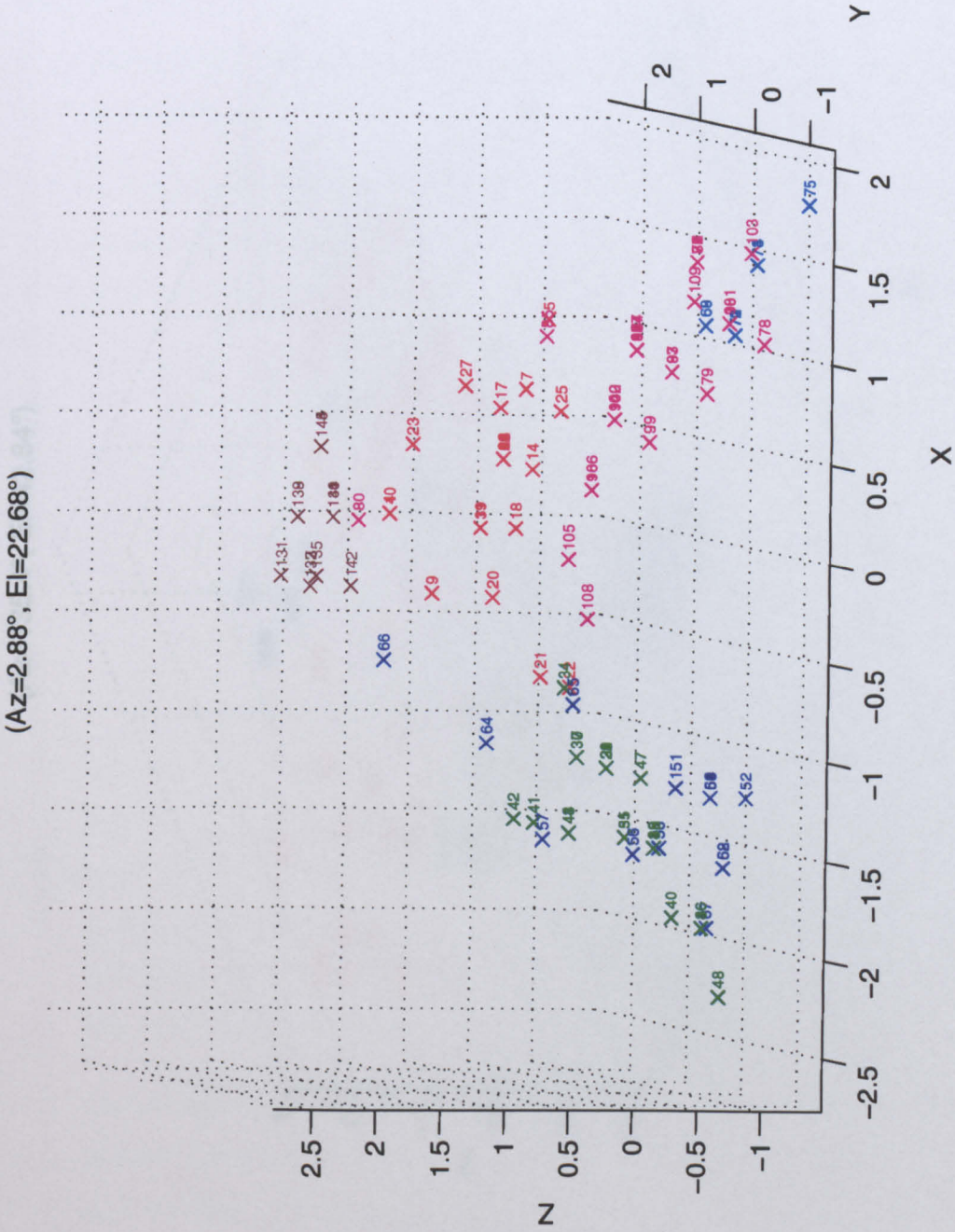


Figure 3.12: Labelled Plot of Factor Scores for Individual Stimuli Resulting from PCA of the Responses for Subject 6 (Musician)

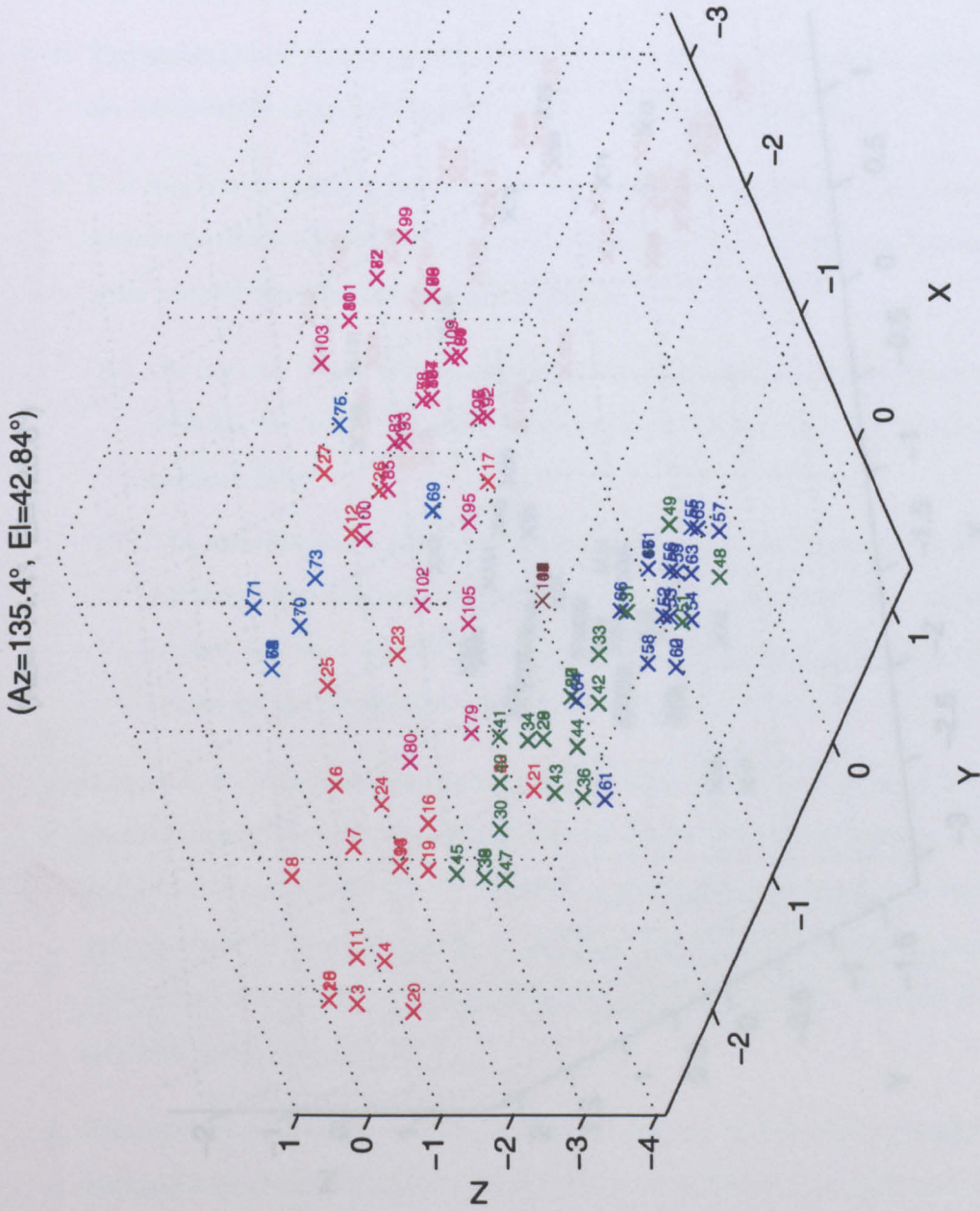


Figure 3.13: Labelled Plot of Factor Scores for Individual Stimuli Resulting from PCA of the Responses for Subject 9 (Non-Musician)

(Az=-14.4°, El=125.6°)

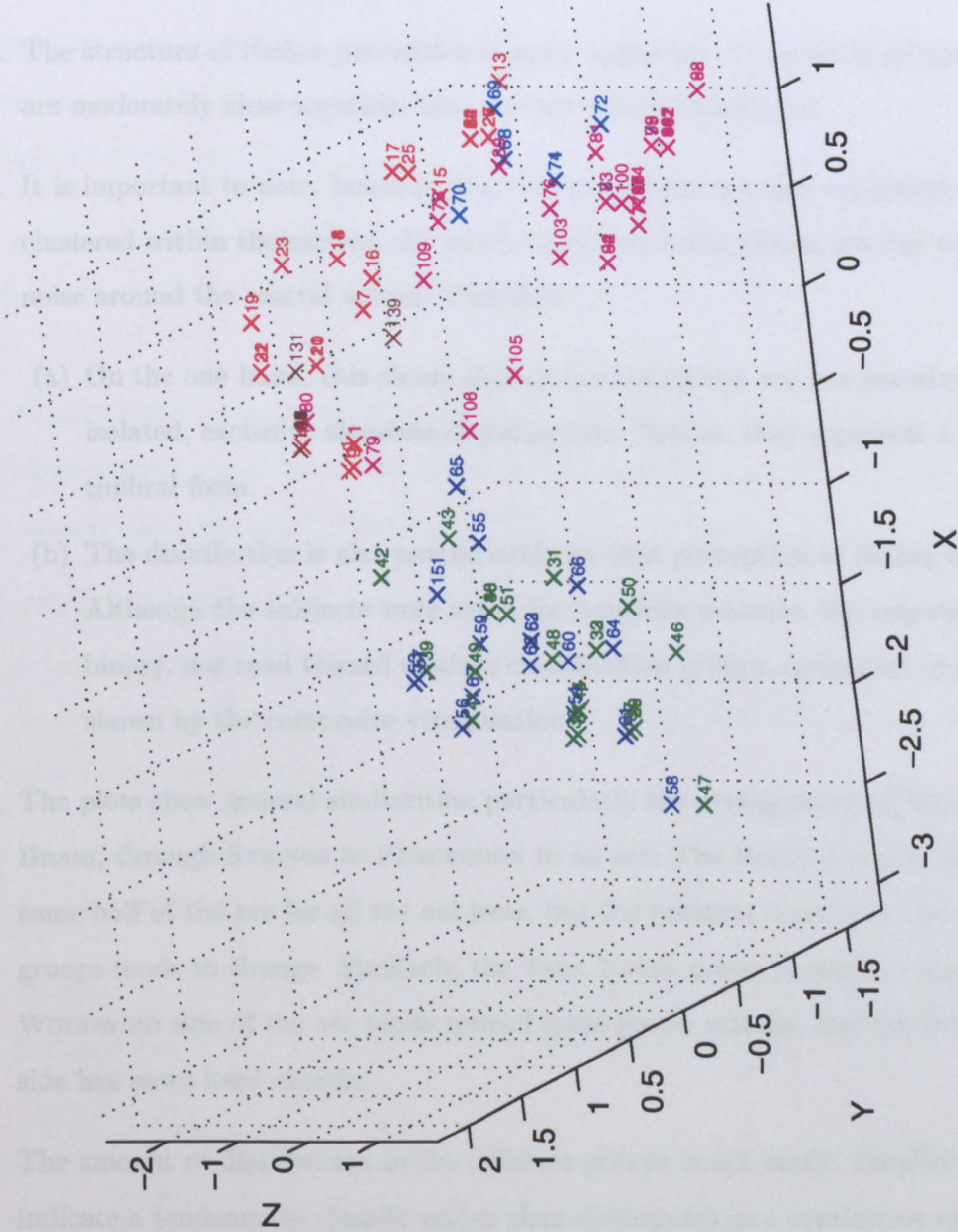


Figure 3.14: Labelled Plot of Factor Scores for Individual Stimuli Resulting from PCA of the Responses for Subject 12 (Non-Musician)

Figure 3.10 shows the colour-coded arrangement of the stimulus groupings. The large dot plots convey the structure of the space well. Although less precise than small markers, the large dots give a better impression of depth and spread. Note that multiple points in the same positions exist, however. The author interprets the plots as follows:

1. The structure of timbre perception is again apparent. The stimuli grouped by colour are moderately close together; they are not wildly distributed.
2. It is important to note, however, that the groups are not well separated, nor tightly clustered within themselves. As will be discussed later, this is not due to there being noise around the central values. Therefore:
 - (a) On the one hand, this shows that orchestral groups are not perceived as being isolated, exclusive elements of perception. Rather, they represent a range of timbral form.
 - (b) The distribution is also partial evidence that perception of timbre is continuous. Although the subjects were asked for template matches, the responses are not binary, nor tend toward distinct classification groups, orchestral or otherwise, as shown by the composite visualisation.
3. The plots show general similarities; particularly the arrangement of WOODWIND and BRASS, through STRINGS to PERCUSSION in an arc. The PIANO group is located on the same half of the arc for all the subjects, but the position relative to the neighbouring groups tends to change. Similarly, the TEST TONES group position is variable. The WOODWIND side of the arc tends toward more gentle attacks, and the PERCUSSION side has more hard attacks.
4. The amount of distribution in the different groups is not static. Smaller groups may indicate a tendency to classify rather than distinguish in a continuous manner, or a lack of ability to distinguish as effectively between stimuli. On the other hand, greater distribution may indicate more difficulty in forming a consistent perceptual image of the template being matched. Similar arguments can be applied to the amount of overlap between the groups. It is impossible to say what exactly processes are at work, but it is apparent that both overlap and distribution seem slightly larger for the non-musicians.

The labelled plots of Figures 3.11 to 3.14 show more detail as regards the relative positioning of individual stimuli. Where a number of items are coincident, the stimulus codes overwrite each other, but the interest value lies in those items that do *not* cluster at the same point. The most important aspect of the plots is that the stimuli have a similar relationship for all subjects. The implication is that the distribution of points is not governed by noise in the data disturbing points around their “natural” values, but that the perception of timbre is structured and continuous in nature. This also relates to the findings of Subsection 3.5.3.

This continuous, structured form is demonstrated by the positioning and type of stimuli which are located between the traditional groupings. For example:

1. Stimulus 105 (cabasa roll) does not fit nicely into the type of form epitomised by those stimuli which form the far end of the PERCUSSION part of the arcs, and so tends toward the centre part of the arc forms.
2. Stimulus 80 (bowed vibraphone) is traditionally a PERCUSSION instrument, but is being played with a bow. This makes it sound STRING-like. Also, its unusual sound makes it somewhat Synthetic-sounding. Therefore it is positioned near the STRINGS and TEST TONES groups.
3. The pizzicato/plucked strings sounds have Percussive aspects, which makes their positions tend toward the PERCUSSION part of the forms.
4. Stimulus 21 (stopped, muted double bass) has a particularly soft sound which places it nearer the WOODWIND part than the other side of the arc.

It is important to remember when considering these plots, that they do not represent 100% of the variance within the 6 axis systems from which they were derived with PCA. As such, the arrangements might be more similar than is apparent when viewing the data in 3D, or might vary in different ways. 3 dimensions is enough to achieve the essence of the form, however.

3.6 Limitations of the Experiment

3.6.1 Limitations of Study Technique

This subsection is concerned with the limitations of the technique used in this perceptual study.

1. Those persons unused to timbral similarity concepts might have found the initial stage of the tests more simple, had the first sample for comparison been reassuringly like the template under consideration. Alternatively, the provision of examples before the tests began could have been used.
2. Familiarity with some sounds and template groups made low-level cognitive judgements hard to achieve at times.
3. The tests were not repeated by the participants, meaning that their consistency could not be formally assessed.
4. During the tests, recent judgements might affect the current one under consideration, but this effect was not assessed.
5. The complexity of the tests was perceived to vary. Strings, Woodwind and Brass were generally perceived to be more easy tests to perform than the Hammered, Percussive, and Synthetic tests.
6. Different groups of stimuli could have been considered, to investigate the effects of the stimulus universe on the discrimination between the stimuli (Subsection 3.3.4).
7. Different sorts of tests might have been useful for comparison purposes and extending the available information. For example, considering more psychoacoustically relevant groups, such as idiophones, membranophones, chordophones, aerophones and electrophones ([22]). Alternatively, semantic descriptions like “harshness” and “hardness” would be of interest.
8. Although the stimuli were presented in a pseudo-random order, having a different irregular order for each participant might have been interesting, to allow analysis of the effects of ordering ([198], [201]).

9. Due to the length of the tests, the template sounds were not repeated before every stimulus judgement. As such, the mental image of the template may have distorted between repetitions.
10. A larger number and range of participants could have been used.

That the technique does in fact work very effectively puts the above limitations into context.

3.6.2 Limitations of the Statistical Methods Employed

The techniques used in this study are not the only way of recording and analysing the information. The statistical methods used are intended to demonstrate that similar conclusions can be drawn from different perspectives, and a coherent picture is apparent. This subsection considers the limitations of statistical methods that have been used.

1. The statistical results are principally limited in accuracy by the accuracy of the information that was gathered. Due to the moderately large number of stimuli, the number of response choices given to the participants was 4. The effect is to maintain consistency through the tests, which might have been compromised by a larger number of choices. However, a by-product of that method is that the variables used in the analysis are more heavily quantised than they might otherwise be. The result of which is that stimuli, which on more numerous scales might be recorded as different, could appear very similar.
2. From another perspective, despite having only a 4 point scale, the participants might have recorded "incorrect" results, which are hidden in the data. This might be due to such factors as momentary distraction or fatigue, resulting in an evaluation different from that which would normally be recorded by that participant. This also adds noise to the analysis.
3. Overall, it should be remembered that the consideration of structure in the test results is not based on direct evaluation of perceived similarity between each stimulus and every other. The solution space in the PCA plots represents a relationship derived from a particular way of recording data. For example, if a piano

sound comes very close to a violin sound, that does not mean that they are perceived as being very similar sounds in all respects, but that within the confines of this study they have *some* aspects which are perceived as similar in the template matching tests which have been performed.

Again, the consistency and strength of the results puts these concerns into context.

3.7 Conclusions

Comparing with the studies listed in Subsection 2.8.2 and other relevant material, the work outlined in this chapter is novel in the following ways:

1. This chapter contains an overview of the processes of perception applied to the study of timbre, of which no other to this depth is known to the author.
2. The perceptual study uses the largest number and range of stimuli to date of a timbre perception experiment, and thus can claim to be more comprehensive and universally applicable than previous attempts studying the same sorts of sounds.
3. The study uses the largest range of statistical metrics of any timbre perception experiment to date in order to reinforce the conclusions from several perspectives.
4. The study successfully pioneers the use of a template matching technique in judging timbral similarity to avoid the effects of non-timbral characteristics impinging on the judgements.
5. The study draws the most wide ranging conclusions to date (as described below).

With reference to the original aims of this chapter outlined in Section 3.2, this perceptual study demonstrates the following points:

1. Timbre space and timbral relationships are not only an engineering model, but also a psychoacoustic one. This is shown by the combination of statistical techniques in Section 3.5. These demonstrate structure in, and a continuum of, timbral perception. The coherent structure of perception in this way shows that a timbre space could be a realistic model of mental organisation.

2. Knowledge of timbral relationships is partly independent of musical training, and is a natural part of perception. The consistency between subjects found in the statistical methods of Section 3.5 is a strong feature. The strength of the correlations for different subjects performing the same tests (Subsection 3.5.2) and the similarity of different subjects' factor score plots in Subsection 3.5.5 are examples of this. Within the scope of this study, large differences between subjects' responses due to musical training have not been found, although more complex analysis might find some.
3. Relationships between timbres are perceived in a logical, structured manner. On a broad level, it is apparent from the correlation coefficients of Subsection 3.5.2 that the relationship between the test results, and thus the perception of the stimuli is consistent across subjects. Similar information is apparent from the variable loadings of Subsection 3.5.4. On a more fine scale coherent structure is shown in the modal results of Subsection 3.5.3 and the factor score plots of Subsection 3.5.5. Groups of sounds are linked together in patterns that are as the author might have expected before the data was gathered; for example, the arc of types from WOODWIND and BRASS, through STRINGS to PERCUSSION in Subsection 3.5.5.
4. Perception of timbre is continuous, rather than categorical. The full nature of the categorical and continuous aspects of timbre perception is not well understood. There are hints in the data that subjects may have a slight tendency to produce categorical or binary (match/no match) answers (Subsection 3.5.1), but in general the evidence is that perception is founded on a structured continuum (Subsections 3.5.5 and 3.5.3). This may well be interpreted at an higher level of perception by decision making processes as described in Subsection 3.3.3, however. It is also worth noting that the continuum results in traditional orchestral instrument group sounds overlapping in perception.
5. Instruments' timbre spaces are not neatly separated in perception, but have intersecting characteristics. This is the most difficult concept of those considered in this study to prove conclusively, because there are only a moderately small number of stimuli compared to those which are audible and only a limited number of which are from the same instruments. Considering the detail of the labelled factor score plots in Subsection 3.5.5 reveals that different stimuli from the same source do produce a considerable range of overlapping positions in the scaled solution space.

Similar conclusions can be drawn from the data of Subsection 3.5.3.

6. A mental timbre space describing a range of timbral form, as imparted from the sonic information in a set of template sounds, can be used successfully to match the features of stimuli which are not part of the template. This, then, matches a wide range of features and avoids the effects of pitch, loudness and duration on the timbre comparisons. That is, the technique used in this study works as hoped. In particular, this is demonstrated by the results of Subsection 3.5.3.

Having established these points it is then desirable to associate these perceptual results with acoustical forms, as shown in later chapters.

Analysis-Synthesis Model

4.1 Introduction

The analysis-synthesis system described in this chapter is used to generate time-varying frequency spectrum representations of the stimulus set (Appendix A) which are used in Chapter 5 for timbral feature extraction and analysis. This chapter discusses:

1. The concepts behind time-varying frequency spectrum analysis-synthesis systems, and an overview of previous methods employed.
2. The analysis-synthesis system used in this research and its adaptive qualities.

4.2 Background to Spectral Systems

Time-varying frequency spectrum techniques have been used extensively in music and auditory research for a considerable length of time. There are a number of reasons why research concerning timbre often perpetuates the use of the spectral form:

1. The greater proportion of previous investigations into timbre have occurred in the spectral domain. This gives a useful starting point for further work.
2. The human hearing system is partly orientated around the perception of sound in frequency bands.
3. Timbral perception can be related to aspects of time-varying frequency spectra (see Sections 2.9 and 2.10).
4. The spectral form is very convenient in its manner of distributing a complex acoustic form for purposes of investigation.
5. The spectral form is not limited by the boundaries of that which is possible to describe accurately in terms of a physical system, such as resonating solids, movement of liquids or aerodynamics ([67]).
6. There are a considerable number of different techniques which can be used to achieve a spectral result with different resulting information representations.

A “plain” spectral form can be considered to be the result of a simple transformation from the time-domain representation of an acoustic wave into a set of time-varying frequency bands (“bins”), such as that which results from a Discrete Fourier Transform (DFT). A fundamental aspect of such a form is that it neither relates directly to the structure of auditory perception, nor the natural processes of sound production to be found in nature. As regards the human auditory system, the *peripheral* hearing system (Figure 4.1) has a structure which relates to the spectral form. This results from the way that the cochlea of the ear transforms time-varying motion of the ossicles into excitation of different nerves along the basilar membrane depending upon frequency ([42], [88]). The biological purpose of this is to facilitate the conveyance of information from the periphery to the auditory cortex with minimum loss of content ([229]). That is, by splitting up the

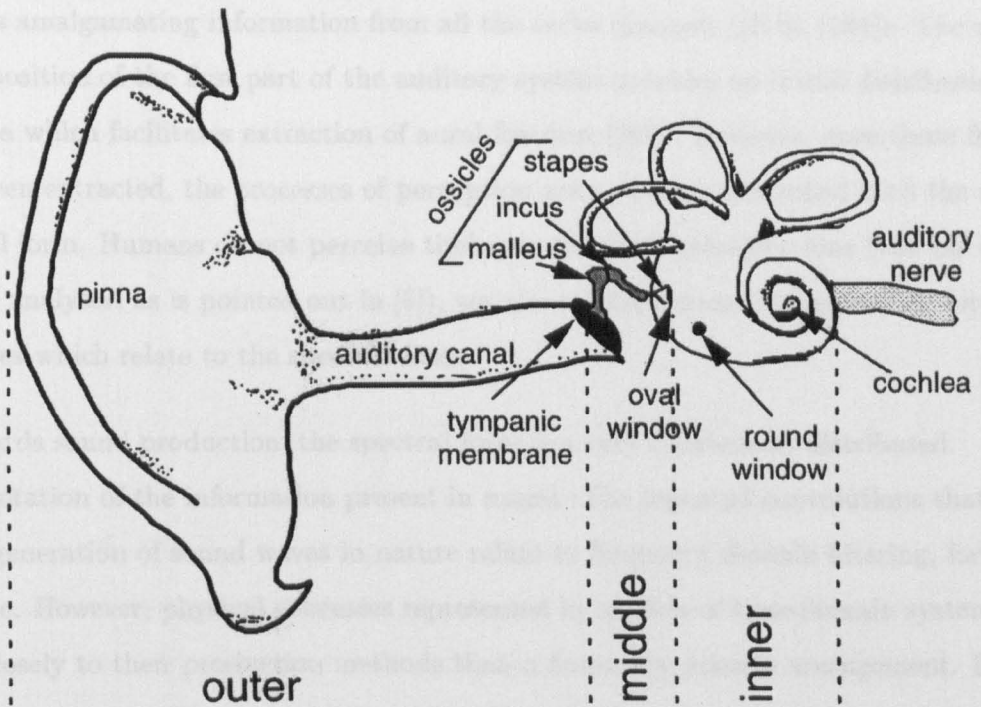


Figure 4.1: The Human Peripheral Hearing System

frequency range of the auditory signal it is possible to convey more bandwidth to the auditory cortex than would be possible with a single neural fibre.

The effect of this method of frequency decomposition is that there are certain distortions imparted to the information which arrives at the outer ear before it reaches the cortex. In particular, multiple frequency components which fall in the same critical band¹ cannot be perceptually separated and generate roughness and beats ([88]). There are other, more complex frequency effects that occur in the periphery and the cortex, such as active cross-channel masking and non-uniform frequency shaping ([223], [78], [134]). Additionally, the task is not aided by the necessity for the hearing system to cope with an extreme pressure range, binaural effects and other problems such as ear wax which add extra distortions which the cortex must deal with.

Although the aural system has some prominent spectral aspects, it does not necessarily follow that the structure of timbre perception is based on the raw spectral form as described by a set of frequency bins which are roughly independent in perception. An example is the general spectral envelope, which has a significant role in perception, yet

¹basic “filter bandwidth” of the ear

involves amalgamating information from all the nerve channels ([210], [184]). The spectral decomposition of the first part of the auditory system provides an initial distribution of the data which facilitates extraction of aural features ([89]). However, once those features have been extracted, the processes of perception are no longer concerned with the raw spectral form. Humans do not perceive timbre in terms of frequency bins (the ear is not a Fourier analyser, as is pointed out in [3]), yet the auditory system is limited by biological processes which relate to the spectral form.

As regards sound production, the spectral form is a very convenient, distributed representation of the information present in sound. The repeated convolutions that occur in the generation of sound waves in nature relate to frequency domain filtering, for example. However, physical processes represented by models of time-domain systems relate more closely to their production methods than a frequency domain arrangement. It is also important to remember that a time-varying frequency spectrum is not a complete description of the sound source from which it was produced, but rather that of a single *instance* of sound production from the source.

Overall, the spectral form is a convenient representation for analysis, modifications and synthesis. This is emphasised by the fact that many researchers have employed spectral techniques that relate neither directly to sound production nor to the ear. There are examples of systems which are less arbitrary about the transform method employed, particularly as regards models of the peripheral hearing system ([201], [200]). Most importantly, however, spectral analysis-synthesis techniques have been shown to be adequate for achieving representations which can be related to timbral forms in a number of situations, without needing to be good models of particular natural processes. It is of more interest to know what the *central* auditory cortex does with the time-varying frequency domain information. That is, spectral analysis is a tool for generating a form which facilitates the extraction of timbrally relevant information pertinent to perception at a later stage.

4.3 Previous Time-Varying Frequency Spectrum Systems

This section presents a brief overview of the sort of spectral systems which have been used to consider sound qualities in the past. So far, the “spectral form” has been considered as

if it means a relatively specific transformed version of the time-domain representation of the acoustic wave. In fact, many different methods exist which have different ways of approaching such aspects as the following:

1. The shape of the filter responses. The shape of the filters may be specifically tailored to achieve results of a particular style. For example, GammaTone filters mimic those of the basilar membrane ([17]), and so facilitate visualisation of the sort of filtering that occurs in the peripheral hearing system.
2. The size and number of filters. In the Discrete Fourier Transform technique the bandwidth of filters is uniform with frequency. In other techniques, such as the constant-Q transform ([18]), the intention is to have filters which approximate the size of the critical bands of the ear, which increase with frequency ([135]).
3. Interactions. Effects such as the cross-filter interactions in the human hearing system can be modelled by analysis systems.
4. Information extraction. Different methods exist for extracting data forms which can be related later to timbral perception; such as finding frequency tracks, formants or noise bands.

This research investigates a large range of sound types. A considerable amount of the analysis-synthesis literature is specific to speech; such as [144], [2] and [124], as well as the large number of books on speech processing ([170], [227], [16], [55], [86] and others). Speech analysis-synthesis methods are not always directly applicable to all types of sound. More generalised methods have been developed such as those detailed in [195], [85], [181] and [57] which do not assume a particular source “instrument”, but are often orientated toward musical, rather than “everyday” sounds (Subsection 3.3.2).

The Discrete Fourier Transform (DFT) and the more efficient equivalent form the Fast Fourier Transform (FFT) have been used extensively in previous research for transformation between the time and frequency domains. They are simple methods which have not been tailored to any particular application and are well documented (such as in [37] and [90]). Their problems as regards sound analysis is the uniform bin size (unlike the ear) and problems with accuracy when used as a Short-Time Fourier Transform (STFT); that is, using a moving window rather than transforming the whole sound at once (for

which windowing the data is required, see [81]). However, the STFT (STDFT / STFFT) is the technique of choice in many modern systems ([124], [195], [27], [182], [181], [39], [85], [58], [57], [95], [183]).

A considerable number of performance-enhancing variations on the DFT family exist ([9], [114]). However, some authors have considered more fundamentally different ways of tackling the problems associated with DFTs. A DFT is a uniform filterbank, therefore filterbanks are a reasonable alternative which provide the ability to tailor the response of the transform more appropriately to the response shape of the ear ([125]). Typical techniques are detailed in [207], [208], [145], [169], [48], [101], [38], [49], [104], and [18].

Once a spectral transformation has been achieved, it is then necessary to extract information of relevance to the investigation. The author calls the result of this the “manipulable form” as it allows analysis and manipulation of data with more direct relevance to timbre, rather than the raw spectral bins. In this research, the extraction of partial tracks is of interest, which can be related to timbral features, such as those described in Section 2.10. Those tracks are the internal components of the sound. They are not static frequency elements, but vary in position and amplitude over time. Methods of tracking partials are described in such articles as [124], [195], and [181].

The synthesis scheme which is used to construct a time-domain result from the manipulable form can be of many sorts, depending upon how the information is considered after analysis. If it is still in a partial tracks form then the additive synthesis procedure is most common (for example [181], [85] and [52]). The precision of additive synthesis may not be required when dealing with larger bands of frequency data however, where inverse transforms ([181]), subtractive synthesis ([21]), amplitude modulation ([21]) are some typical alternatives.

Although the basic concepts of analysis and synthesis to and from the spectral domain are clear from a general perspective, the finer details are often more complex. This is indicated by the fact that some techniques rely on an amount of user configuration to tailor the system appropriately to the type of sounds being considered ([182], [181], [85]). This is because systems are often unable to adapt their approach automatically to the type of sound, as the ear does. Also, techniques are employed to simplify the task of analysis by taking advantage of the fact that the system need not necessarily be ear-like and real-time.

For example, by starting analysis in a region where the sound is well established and working backwards through the sound ([85], [181], [21]), it is sometimes more easy to establish partial tracks than analysing from the start of the sound.

4.4 Analysis-Synthesis Technique Overview

The purposes of the spectral analysis-synthesis system used in this research are as follows:

1. To convert the time-domain representations of the input stimuli (Appendix A) to time-varying frequency-domain distributed representations, which can be used for the extraction of spectral features, which in turn can be analysed for importance in distinguishing the timbral qualities of sounds (Chapter 5).
2. To investigate the problems associated with a system which adapts its time-frequency analysis viewpoint depending upon the nature of the sound.

The first purpose above is a rough analogue of the sort of processes that might take place in the human hearing system, as described in Section 3.3. This is necessary to facilitate the investigation of timbral form based on acoustical information in Chapter 5. It is not, however, intended to closely mimic the actual physiological anatomy of the human auditory system, which is an exceedingly complex and little understood area. It is, rather, an engineering tool for allowing investigation of timbral structures. The second purpose is a more exploratory area of consideration concerning how to model the adaptive qualities of the human hearing system:

“A major technical problem for the analyst is that the brain can perform simultaneous time and frequency analysis, that is, it can operate in both series and parallel modes simultaneously.” [161]

“[It is hypothesised] . . . that a given spectral component will be simultaneously processed by analyzers with a variety of bandwidths centred on that component. The result would be equivalent to a “multiple-bandwidth spectral analyzer,” which could be of particular value in the discrimination of natural sounds that differ in spectral shape, tilt, or contrast.” [150]

The auditory cortex appears to consider sonic information from several different viewpoints in order to adapt perception to an appropriate level of detail depending upon the content of the sound. There is necessarily a time-frequency trade-off as defined by Gabor's uncertainty relation, also known as the Schwarz inequality ([65]):

$$\Delta t \Delta f \geq 1 \quad (4.1)$$

Therefore, in a single transformation from the time-domain to the time-frequency domain, the time-resolution of the information must decrease as the number of frequency bands increases. This is neatly demonstrated by the equation for the Discrete Fourier Transform (DFT):

$$X(k) = \sum_{n=0}^{N-1} x(nT) e^{-jk\Omega nT}, k = 0, 1, \dots, N - 1 \quad (4.2)$$

Equation 4.2 shows that as N increases, the number of result bins (X , indexed by k) increases, but necessarily the sum uses more input time-domain samples (x , indexed by n). As such, frequency resolution trades against time resolution. If the ear is viewed as a bandwidth-limited spectral analyser in the most loose sense, it also can be seen to be limited by this relationship. Within the time information are considerable cues as to the nature of sonic events through temporal proximity; how events are initiated, where events are occurring (through inter-aural time displacement) and so forth. Frequency information allows the discrimination of different pitches, resonances of the sound source, the precise relationship between partials and so the nature of the physical materials which caused the event, and so on.

It is apparent from previous studies that humans hear events with high precision in both time and frequency. For example, a recent study of general temporal acuity produced an estimate of about 2ms ([47]). Yet, humans can also perceive differences in frequency to less than 0.3% in the range 0.5-2kHz ([54]). To accommodate this phenomenon, visualisation of phenomena such as speech is best accomplished by the use of separate plots; one with greater emphasis toward time resolution and another with greater frequency resolution.

Although time-frequency resolution is a very complex area, the conclusion that can be drawn is that human hearing has evolved to extract details in a manner that befits the

condition of the aural World at any particular moment ([47]). That is, perception is adaptive (as noted previously in Subsection 3.3.1). For example, the beginning of aural events tends to be dominated by the sonic transients generated by a force causing the event to occur; such as the scrape of a bow on a violin string before the string settles to a resonating pattern. The ear is not interested in the long-term content at that point, such as the pitch or the resonant structure, as they have not yet developed. It is desirable to diagnose the cause of the event. Therefore, time resolution is most important in that situation. But if the source is in a less transient condition, then the content will be assessed by the ear more in terms of the frequency relationships that exist between the partials.

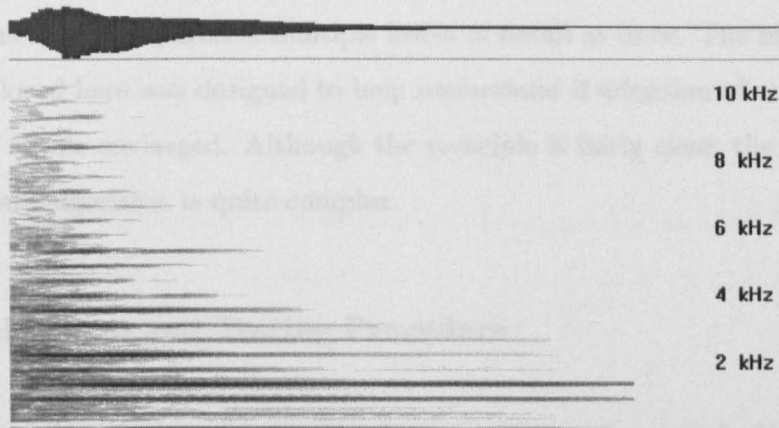


Figure 4.2: 512 point FFT Spectrogram Underneath Time Domain Form of Stimulus 8 (martele violin)

An example of the changing nature of the acoustic form of natural sounds with time is shown in the FFT of Figure 4.2. The transition is apparent from high energy and high frequency spread at the onset as the bow starts to drive the string, to the resonating periodic pattern later in the sound. The ear's interpretation of such an acoustic form is a complex topic. However, the system used in this research tests the basic concept of an adapting analysis for producing spectra which can be used as data in later stages of the research (Chapter 5).

4.5 Method and Results

The reason for the analysis stage is to translate the time-domain samples of the stimuli used in this research into partial track representations. Those tracks are the amplitude and

frequency trajectories of partials which have been extracted from the results of the transformation of the samples to time-varying sets of frequency domain bins. Those sets of bins have different time and frequency resolutions to allow adaptive selection of the most appropriate resolution in different parts of the spectrum depending upon the prevailing “stability” of the data.

The methods used in this research are largely adapted from other schemes, apart from the stability concepts as detailed in Subsection 4.5.2, and their application. However, other authors have considered the concepts of multiple level analysis before ([39], [166]). This has similarities with the effect of the human hearing system’s ability to analyse sound, coping with a necessity to perceive multiple levels of detail at once. The relatively simple technique employed here was designed to help understand if adaption of sonic viewpoint has the sort of effects envisaged. Although the principle is fairly clear, the implementation of an appropriate algorithm is quite complex.

4.5.1 Development and Testing Procedure

The analysis tool described in this chapter was refined over the period of its development. The aim was to achieve results synthesised from the analysed time-varying spectral forms which matched the original time domain forms as closely as possible. That is such that the time-varying spectral form is suitable for use in the timbral feature extraction and analysis detailed in Chapter 5. The development process was not a trivial task, due to the complex adapting nature of the analysis combining different viewpoints on the data. In addition, the analysed form is not a simple transformation of the original data. If the analysis-synthesis system was only a DFT followed by an inverse DFT (IDFT) of the entire sound, then the synthesised versions would be the same as the originals. However, the analysis creates a “manipulable” form of partial tracks, extracted from the data, which is generated to facilitate extraction and analysis of timbrally-relevant features. It is from this form that the results are synthesised.

To achieve comparisons between different configurations of the analysis system, a number of testing stimuli were used to examine whether the algorithm adjusted appropriately to the type of stimulus. These were as follows:

1. White noise. This facilitates examination of wideband/noise performance.
2. Sine tone, Square wave. These test the opposite to white noise; that the system focuses on the steady, harmonic details.
3. Pink noise with embedded high amplitude sine tone. This is asking a more complex question of the system; that it can find the steady components accurately even when surrounded by more noisy elements.
4. Rising klaxon note. This checks that highly mobile frequency components can be tracked.
5. Accordion, Celesta, Harmonica tones. These examine more “real-World” spectral forms both in frequency (wideband and harmonic forms) and time-varying performance.
6. Vocal phoneme /IY/ (“ee”). Although beyond the specification of the sounds used in other parts of this work, this was included as an additional comparative tone, especially for its formants and strong harmonics.

Due to limitations of time, it was not possible to utilise the judgement of subjects other than the author for comparing relative performance of the analysis tool in different configurations, except on an informal basis. The success of the development of the analysis-synthesis tool is thus related to the most effective configuration that the author could achieve in the time available. The author, however, refined the analysis tool in a controlled manner. This involved stepping through the values of configuration parameters for all stages of the process to hand-optimize the values both for the stage in question, and in the context of the overall analysis system.

There were two sources of information used in configuration. Firstly, the resynthesised results provided an acoustical comparison, but also there are graphical outputs from every stage of the analysis-synthesis system. The latter allowed examination of the data forms at different points in the analysis process, and also the functions generated by the system which control the adaptation. The visual feedback was particularly helpful when deliberately probing different parts of the system using the test stimuli described above.

The final part of the development and testing was when all 153 of the sounds used in this research (Appendix A) were analysed. The resynthesised versions were then individually

checked for quality by the author compared to the originals before proceeding to use the data for timbral feature extraction.

4.5.2 Stability Concepts

The ear chooses the appropriate time-frequency resolution for the conditions in the signal. That is, improved time resolution where content is unstable (less periodic), and improved frequency resolution where content is more periodic. There is no point in analysing for frequency relationships in noise, or rapid changes in a perfectly stable waveform. The time-frequency resolution of the analysis necessarily affects how many partials can be extracted, and with what temporal accuracy. This in turn changes the nature of the analysed/re-synthesised result.

The appropriate level of detail can be determined through data “stability”, which can be related to periodicity. Stability in this research is determined from a combination of several metrics. The metrics consider differences in the following aspects:

1. The positions of zero crossings.
2. The gradients at zero crossings.
3. The positions of peaks between the zero crossings.
4. The peak amplitudes between zero crossings.
5. The density of peaks between zero crossings.

An example waveform with marked measurement points is shown in Figure 4.3. No single metric is perfect at measuring periodicity, but a combination of simple metrics can be much more effective. Stability is used in this research to consider both the nature of the complete signal (Subsection 4.5.3) and also in individual frequency bands (Subsection 4.5.5). This allows simultaneous consideration of the “macro” and “micro” aspects of the sound. Furthermore, stability has different values over time, which can be found by considering a moving data window.

The stability is considered to have a certain number of “levels”, where an increase of one level corresponds to a doubling of length of samples over which the waveform is considered

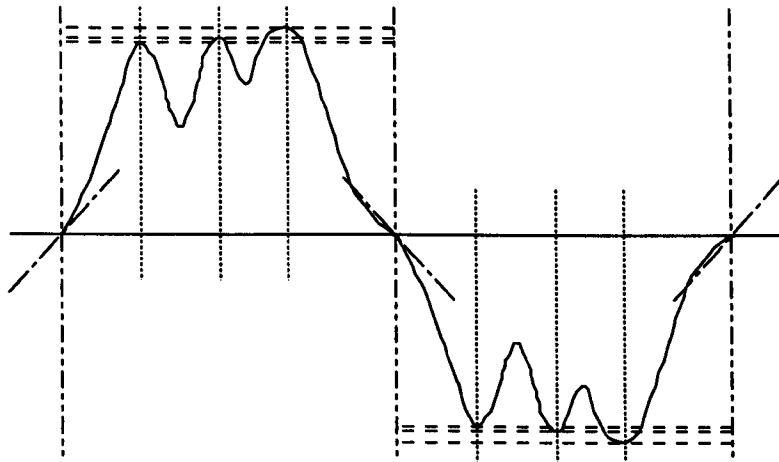


Figure 4.3: Example Waveform with Marked Periodicity/Stability Measurement Points

to have a significant degree of stability. Those levels correspond to increasing frequency resolution of analysis. Individual metrics are measured as:

$$\text{metric result} = \frac{1.0}{\text{number of distinct values in the length under consideration}} \quad (4.3)$$

Distinct values are determined by comparing each item of interest (such as zero crossing gradients) against every other in the length of consideration with the additional effect of a tolerance value. Metrics are combined as a weighted sum, and the final stability amount is the highest level to pass a stability “break point” value.

4.5.3 Part 1 : Analysis of Stability of Input Data

The aim of analysing the stability of the input data before it is filtered is that it provides overall information which can be used in Part 3 (Subsection 4.5.5) to augment the information obtained from filtered regions of the spectrum. It also aids the researcher in understanding something of the general nature of the sound.

It was found that a minimum (level 1) window length of $500\mu\text{s}$, a tolerance of 1% and a stability breakpoint of 30% seemed to act in the desired manner with a range of test sounds. That is, where rapid, noise-like changes occurred, the program chose the lowest level of stability. For more periodic parts, the result level increased, and was highest for strictly stable regions. Variations to window length naturally allow more rapid or slower variation in level, tolerance allows more or less stable regions to appear stable and breakpoint also affects how easily stability is registered. These confirmed the expected

effects. The metric weightings that appeared to offer the most appropriate balance for a range of sounds were as follows; zero crossing positions, 12%; zero crossing gradients, 28%; peak positions, 24%; peak amplitudes, 6%; peak densities, 30%. Many interpretations could be placed on these results; most importantly, a single metric is not adequate.

The results are interesting when displayed graphically. The extremes of stability are apparent in Figures 4.4 and 4.5 (white noise and A5 sine wave respectively). These show the outputs from the author's program for this part of the analysis. The screen dumps show the time waveform, and the located zero and peak positions. Below these are the metric values (greater bar height implying greater likelihood of the waveshape being periodic in the window of consideration) for 5 levels of increasing window length (doubling for each level going down the screen). The lowest plot is the resulting chosen level, which can display 5 levels of bar height corresponding to the greatest level at which the waveform is considered stable. The grey bars display the longer term trend.

The stimuli used in this work often display the sort of change of characteristic through the sound as described in Section 4.4. That is, from the region of transients at the start of the sound to the resonating pattern later on. For example, Figures 4.6 (showing the scrape of the bow at the start), 4.7 (showing the pluck of the string at the beginning), and 4.8 (with the breath of the performer driving the flute into resonance through the sound).

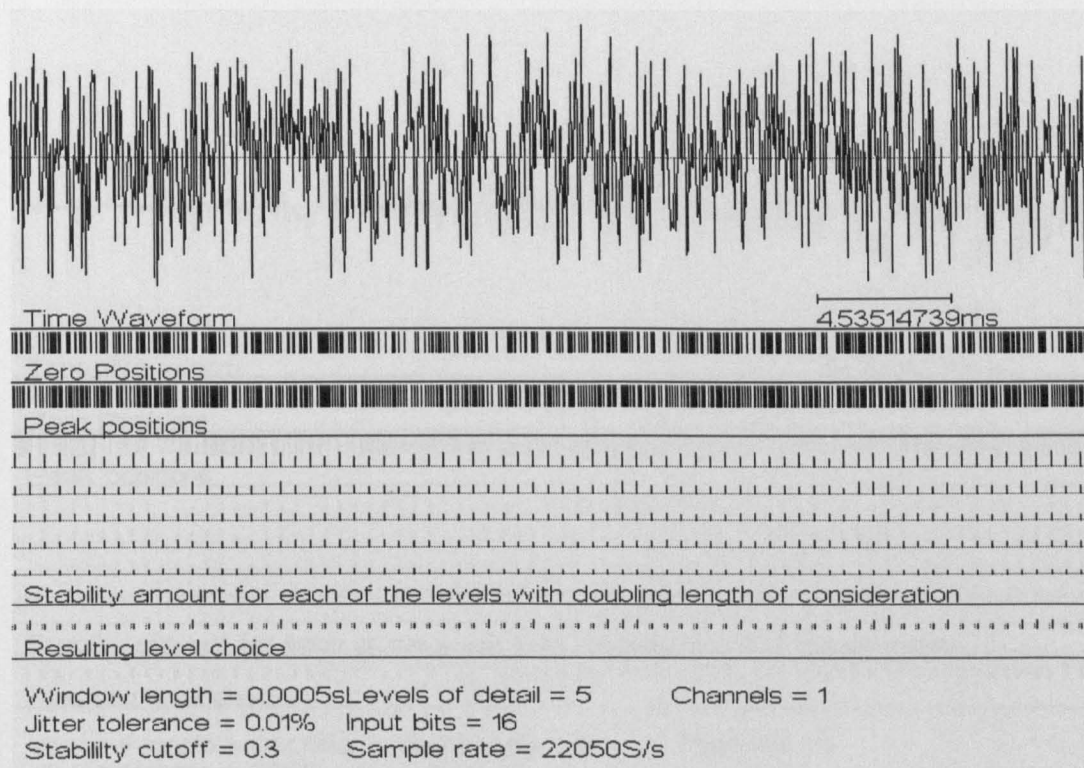


Figure 4.4: Part 1 Stability Analysis Screen Dump; Stimulus 145

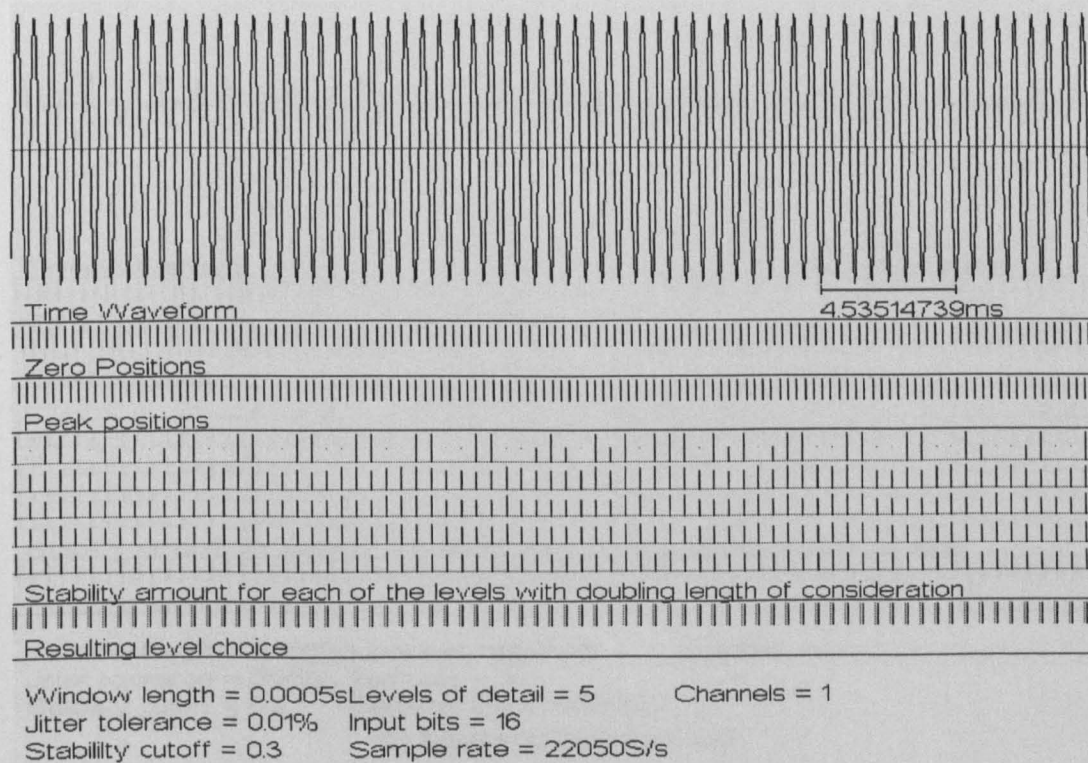


Figure 4.5: Part 1 Stability Analysis Screen Dump; Stimulus 136

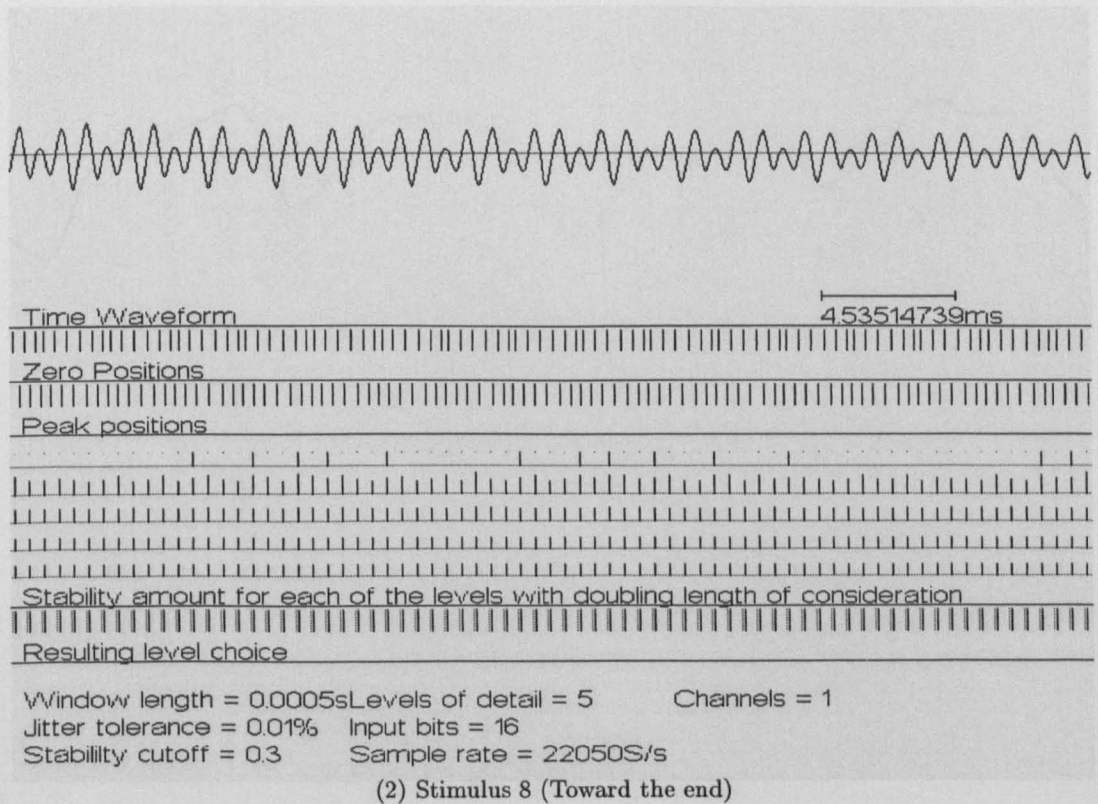
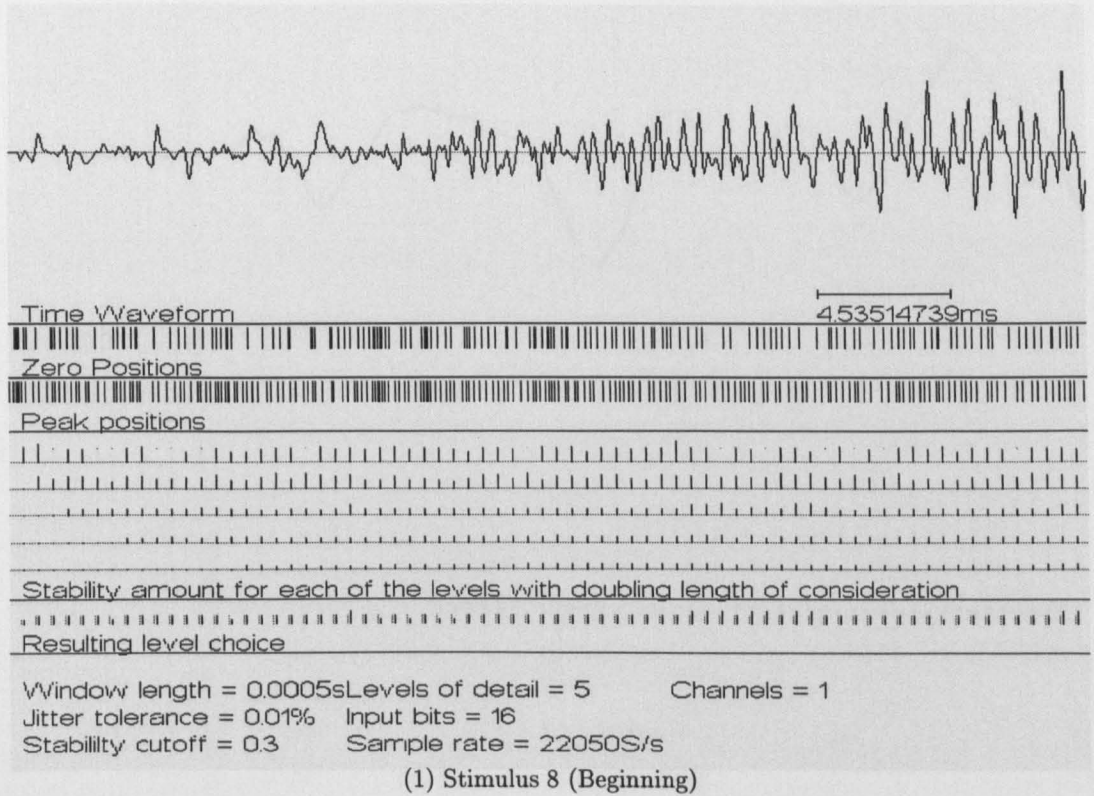


Figure 4.6: Part 1 Stability Analysis Screen Dump; Stimulus 8

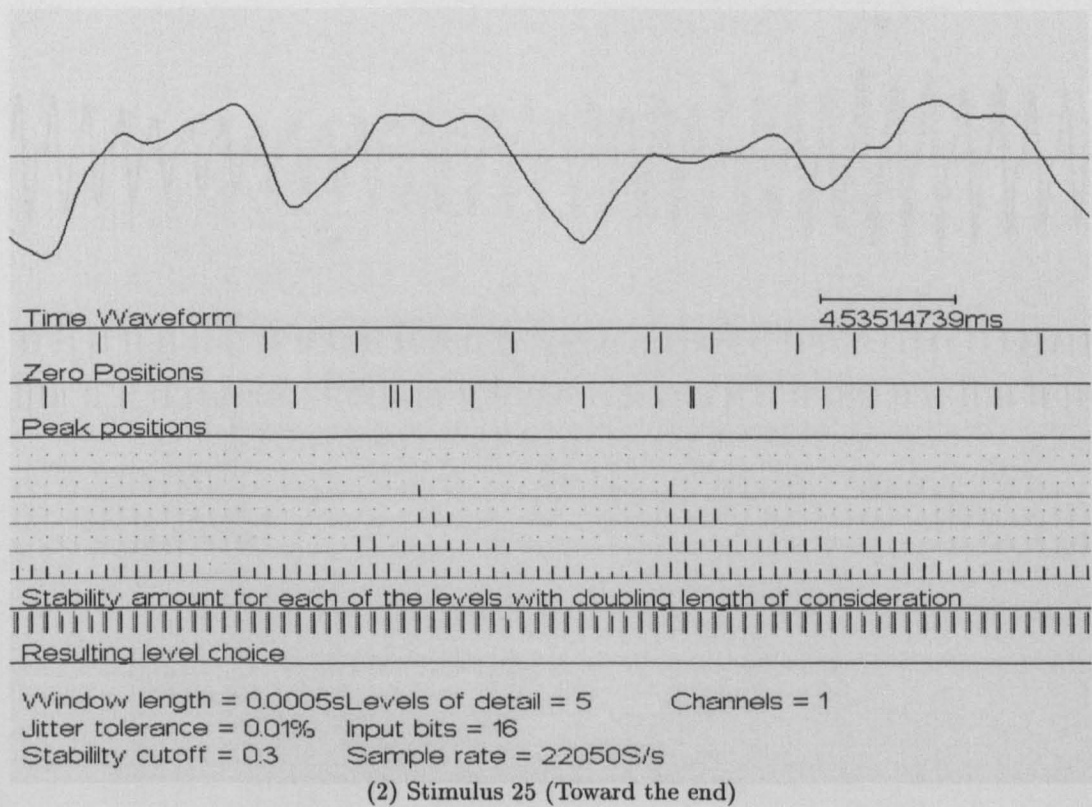
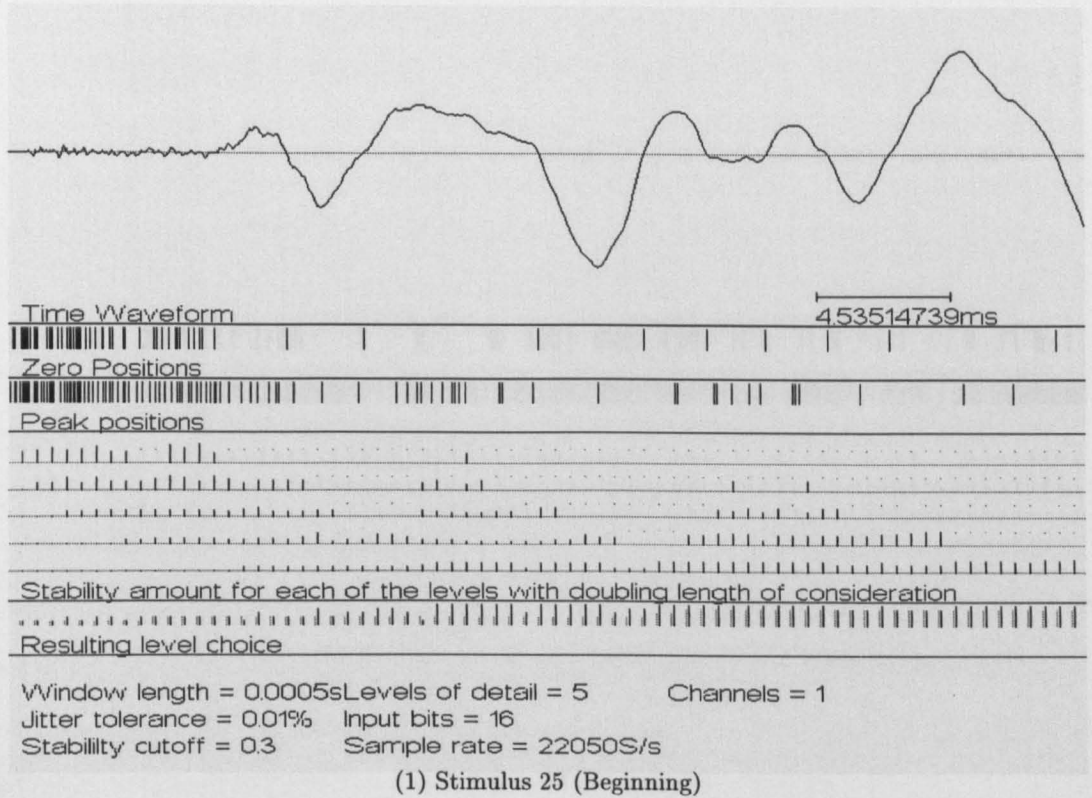


Figure 4.7: Part 1 Stability Analysis Screen Dump; Stimulus 25

4.5.4 Part 2 : Multiple-Resolution Analysis

The goal of this part of the analysis is to generate a sequence of resolution levels across the domain. The method is thus recursive, starting with the highest resolution and progressively decreasing resolution. There are also methods for generating a resolution level that is not necessarily a power of two.

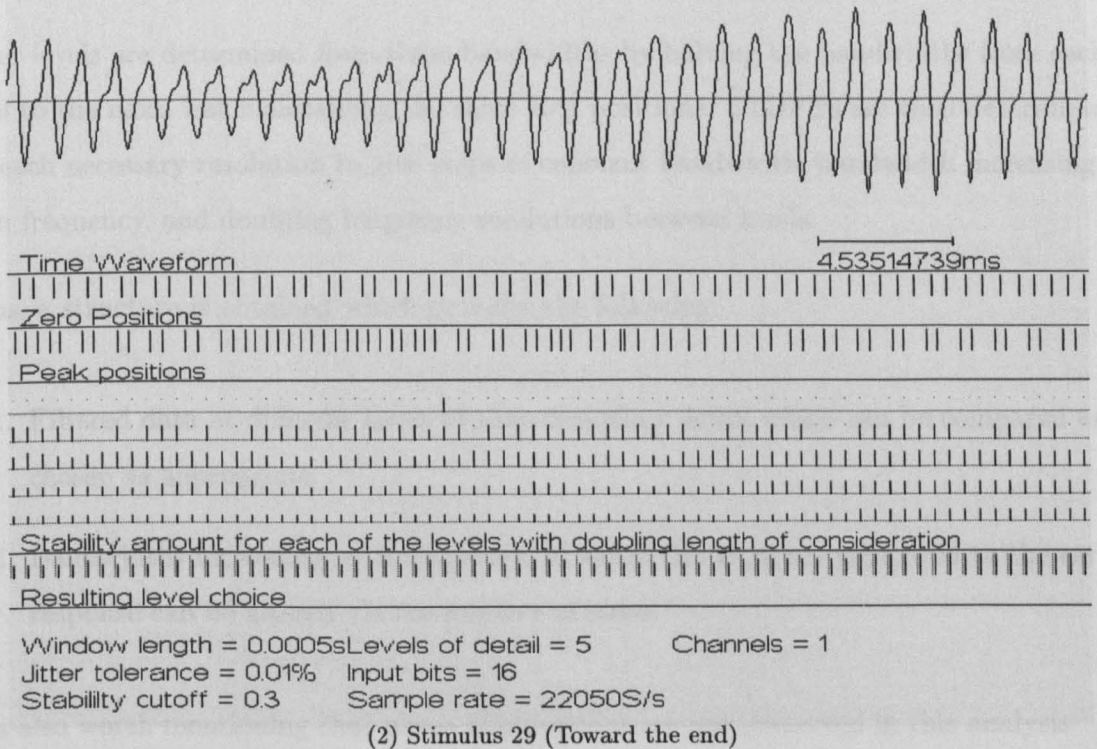
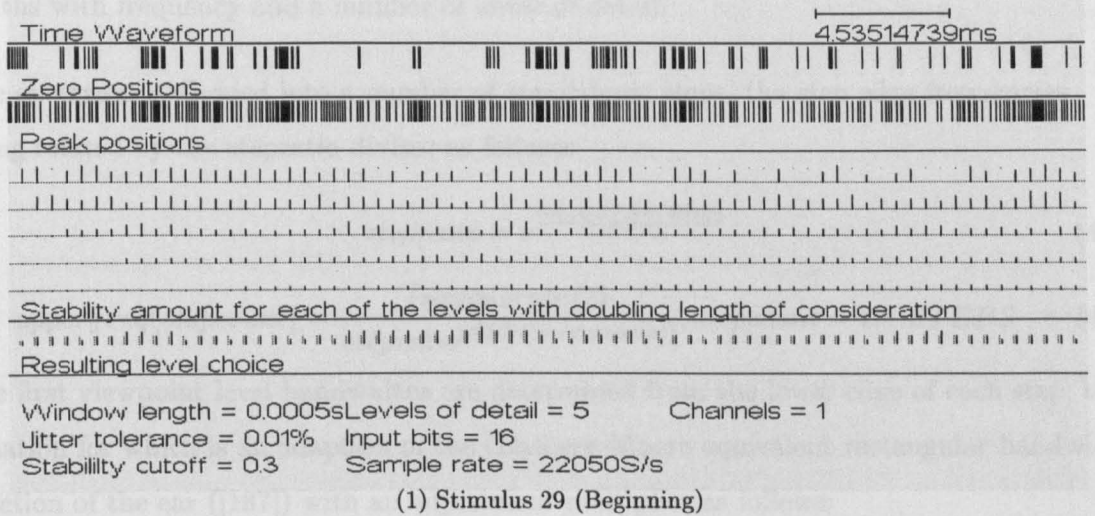


Figure 4.8: Part 1 Stability Analysis Screen Dump; Stimulus 29

4.5.4 Part 2 : Multiple-Resolution Analysis

The aim of this part of the analysis is to generate a number of viewpoints in the spectral domain. The method in this research utilises STDFTs for simplicity, although any spectral transformation method could have been used. A number of STDFTs are taken to produce different resolutions. These are used together to give a structure which has increasing bin widths with frequency and a number of levels of detail.

The spectrum is divided into a number of logarithmic steps, the step edge frequencies being related by the *stepratio* divisor as follows:

$$\text{stepratio} = e^{\frac{\ln(\text{LOGVAR} * 20000)}{\text{STEPS}}} \quad (4.4)$$

$$\text{upper freq}(\text{stepcount}) = \frac{(\text{samplerate}/2)}{\text{stepratio}^{(\text{STEPS} - \text{stepcount})}}, \text{stepcount} = 1 \dots \text{STEPS} \quad (4.5)$$

The first viewpoint level bandwidths are determined from the lower edge of each step; the equation for which is an adaption of the Glasberg-Moore equivalent rectangular bandwidth function of the ear ([187]) with an adjustment multiplier as follows:

$$\text{ERB}(\text{adj}(f)) = \left(\frac{f * \text{EARQDIV}}{1000} + \text{MINBW} \right) * \frac{3.0}{\text{OCTDIV}} \quad (\text{Hz}) \quad (4.6)$$

Finer levels are determined from these bandwidths by halving the bandwidths from each level to the next, but maintaining the same step positions. STDFTs are then determined for each necessary resolution to give steps of constant bandwidth, bandwidth increasing with frequency, and doubling frequency resolutions between levels.

Thus, a structure is obtained which provides the following:

1. Filtered data at different levels of time-frequency detail which can be compared and chosen as appropriate.
2. Bandwidths increasing with frequency, of which the accuracy compared to the ear's response can be altered via the number of steps.

It is also worth mentioning that phase relationships are not preserved in this analysis method. This is done to simplify the amount of data and complexity of the system. Phase is known to have some minor effects on timbre ([156], [157], [173]), particularly with respect to non-static waveshapes.

4 resolution levels and 7 steps, with a *LOGVAR* of 0.01, across the frequency range were found to be adequate for testing the concepts of interest. The levels and steps parameters necessarily affect the subsequent Parts 3-5. It was apparent that both affect the sound quality of the results as expected. In general, increasing the number of steps resulted in the timbre of the resynthesised forms tending toward the originals more precisely. Often this was quite a subtle effect, but shows that matching the analysis to the ear's response does allow partials to be extracted in a more appropriate manner over the entire range of hearing. Using a single analysis bandwidth results in an imbalance in the way that partials are extracted to achieve the manipulable form; and in turn this affects the analysis of timbre through spectral features.

The number of resolution levels is more complex due to the highly coupled way in which that parameter affects the way that many of the other parameters in this analysis scheme must be configured. It was found that having a single level of analysis was detrimental to the resulting sound quality, but necessarily meant balancing the adaption scheme was not the significantly complex problem that it is with more levels. The result is that adaption does make a positive contribution but, as will be described later, it is a complex area. As regards the level 1 bandwidths as obtained from Equation 4.6, the standard Glasberg-Moore values were used for *EARQDIV* (107.939) and *MINBW* (24.7) from [187]. *OCTDIV* was used to vary the first level analysis width, and thus the other levels' finer resolutions. With a fixed number of levels, varying *OCTDIV* necessarily produces a trade-off between the effects of different resolution depths; greater *OCTDIV* providing more precise frequency analysis, but worse larger scale analysis. Ideally, given very large computational resources, large numbers of levels starting at wide resolutions would be used. An *OCTDIV* value of 4.5 was found to produce adequate results.

4.5.5 Part 3 : Analysis of Stability of Bins and Analysis of Amplitudes

This part considers the individual data streams which have been extracted in the basic multiple-level-multiple-resolution analysis and develops the stability and peak amplitude information. The stability analysis is the same basic technique as for Part 1 (Subsection 4.5.3) and thus as described in Subsection 4.5.2. The stability analysis is applied to each of the bins in level 1 (the widest set, as determined from Equation 4.6), rather than the finer frequency analyses of the other levels. This information can then be used in Part 4

(Subsection 4.5.6) to choose an appropriate level of detail in each of the regions associated with the level 1 bins. A difficulty is that different steps have different sized bins and so the amount of bandwidth contained varies. This in turn means that the stability criteria must be different for different steps, as higher frequency bins will necessarily encompass greater waveform variation. The output from Part 1 can be incorporated at this stage to provide overall stability bias as well. This part also finds the peak amplitudes for all bins on all levels over time, which allows amplitude comparisons without the additional complication of periodic time variations. This tactic is used in several systems.

It was found that the different steps of analysis for level 1 required different parameter values to control the extraction of the stability information. To achieve a consistent response to periodicity across the frequency range, balancing parameters as shown in Table 4.1 worked well within the confines of the values for the rest of the system. It is apparent that there is a very uneven look to the parameters; this aids in demonstrating that this is only one of many potential configurations of parameters due to the complex couplings between them.

| Step | Minimum Stability Window (in Samples) | Tolerance (%) | Break point (%) |
|------|--|------------------|--------------------|
| 1 | 1000 | 1 | 0 |
| 2 | 700 | 1 | 30 |
| 3 | 200 | 50 | 30 |
| 4 | 200 | 7.5 | 50 |
| 5 | 300 | 7.5 | 55 |
| 6 | 200 | 10 | 55 |
| 7 | 100 | 10 | 55 |

Table 4.1: Balanced Main Parameters for Part 3 of the Analysis

No obvious relationship was found between the bin sizes and the appropriate parameter values for a balanced stability analysis. One may exist, yet the stepped nature of the system made it confusing, meaning that a hand-configured version performed more effectively. Although in some versions of the parameters coupling the Part 1 output into the model at this stage was found to be useful, in the final version its contribution was not needed. The same balance of weights for the metric results as with Part 1 was used.

4.5.6 Part 4 : Combine Bin Data Based on Stability Information

At this stage, the bin data from Part 3 is translated into peak positions at each timestep. Peaks are found by searching across the frequency range, and at all levels of detail. The data is then reduced by removing peaks that lie beneath perceptual thresholds; that is, those masked by close high amplitude peaks, and peaks of low amplitude compared to the rest of the data (similarly to [58]). Additionally at this stage, the peak amplitudes couple into the stability consideration by providing a small amount of emphasis where peaks exist which are considerably higher than the average. The different levels are then combined into a single representation using the stability data to choose between them.

The adjustable parameters in this part are concerned with the masking effects that occur just before the different levels of resolution are amalgamated. These had a large impact on the size of the representation, without greatly affecting the perceived timbre, which indicates masking mechanisms in the ear. Within steps, peaks were removed which were more than 20dB down on the maximum. Across the spectrum, peaks were removed which were 40dB down on the maximum. As regards peak amplitude to stability emphasis, values greater than 3 times the average cause a 1 level increase at that point in stability for each difference of 3 times. This improves clarity slightly in some circumstances.

4.5.7 Part 5 : Extract Partial Tracks

Part 5 is concerned with extracting partial tracks from the peaks data formed in Part 4. The method attempts to find continuation of peaks between frames, indicating spectral elements' development over time. This matching process considers the three dimensions of frequency, amplitude and bandwidth. These are considered in terms of the current gradient, unless at the beginning of a potential partial track. A bell-like variation profile is used to indicate whether a match is likely, using the form:

$$Matchval(pvalue, multfac) = e^{multfac * pvalue^2}, multfac < 0 \quad (4.7)$$

Sleeping partials are also allowed to prevent the masked partials phenomenon known as "doodley-doo" ([58]).

4.5.8 Synthesis Scheme

Two synthesis techniques were developed to prove the worth of extraction of partial tracks from the representation. The first method takes the results of Part 4 (Subsection 4.5.6) and synthesises, with oscillators assigned to constant frequency positions, from the peak data. The second method uses the extracted partials from Part 5, and assigns oscillators to partial tracks. The latter, then, follows the internal components of the sound, maintaining phase coherence compared to the former, where rapid changes in amplitude are likely to occur as the partials cross through the bin positions. Synthesis also allows examination of the effect of leaving out phase from the analysis-synthesis system.

It is apparent from comparing the two synthesis scheme results (from bins/peaks and partial tracks) that synthesising from partials produces a much better result. The oscillators assigned to partials change their parameters much more smoothly than those assigned to bins. The effect of omitting phase from the representation is to lose a small amount of the timbral character in some instances. The most obvious case is where there is a large and definite low-rate phase shift (say, at sub-20Hz), such as is more associated with a deliberate synthesised effect than orchestral instruments.

4.5.9 General Points

There are a considerable number of potential combinations of parameters controlling the balance of the analysis detailed in this chapter. Much of the investigation process depended upon hand optimisation, by considering each part individually and aiming to understand how the parameters controlled the solution. This was done in a systematic manner, but it was apparent that altering a few key parameters could significantly change the balance of the system. The more general the scheme is supposed to be, the harder it is to balance the system appropriately and with a quick iteration time for all the range of sound qualities of interest.

As regards the overall quality of the system for purposes of providing the manipulable form of the sounds to the subsequent stages, the results can be described as adequate. In particular, the results suffer an amount from switching noise between the levels of detail, particularly for less static sounds, and in other ways due to the complexity of the

implementation. It was not expected that the results would be perfect, due to the nature of the investigation, but rather to show that the adaptive nature of the analysis makes a positive difference to the timbre of the sound which is reconstructed from the manipulable form.

4.6 Conclusions

The time-varying frequency spectrum analysis-synthesis system detailed in this chapter is designed to achieve the following purposes:

1. To convert the time-domain representations of the input stimuli to time-varying frequency-domain distributed representations, which can be used for the extraction of spectral features, which in turn can be analysed for importance in distinguishing between timbral qualities in sounds (Chapter 5).
2. To investigate the problems associated with a system which adapts its time-frequency analysis viewpoint depending upon the nature of the sound.

This work is novel in the following way:

1. The system works with a wide range of sounds without user intervention but has adaptive time-frequency trade-off linked to the periodic/stability conditions within the stimuli without a knowledge-based system with a frequency-domain result. The implementation of this particular combination of features is novel.

The final configuration of the system is not the only possible answer. The complex coupled nature of the system and the large number of parameters means that the arrangement is adequate for the purposes of deriving partial tracks for the next stage of the research (Chapter 5) rather than being a perfect solution. Overall the system represents an interesting way of analysing sound which indicates a potentially more aurally appropriate approach to spectral analysis than single level analysis schemes.

Timbral Feature Extraction and Analysis

5.1 Introduction

This chapter pulls together the major threads of this research, as it investigates the links between perceptual timbre groupings (previously considered in Chapter 3) and acoustical features. A total of 335 acoustical features, extracted from the spectral forms of the 153 sounds of the stimulus set (Appendix A), are analysed. The spectral forms were produced using the system described in Chapter 4. These features relate in part to those found by previous researchers to be important (as detailed in Chapter 2). This chapter discusses:

1. The concepts behind timbral feature extraction and analysis of those features.

2. The particular features considered in this research and the methods used to extract them from the spectral form.
3. Statistical analyses of the extracted feature data and a consideration of the relationship between acoustical data and perception.
4. The structure and dimensionality of timbral perception.

5.2 Background to Feature Extraction and Analysis

“It is not only the question ‘what is a sound made of?’ that we have to answer, but the much harder one of ‘how do we perceive this sound in relation to its constituent elements?’ ... [thus] we shall establish a geography of the sound universe.” [14]

The features which are extracted in this research are designed to facilitate consideration of the types of acoustical information which may enable the auditory cortex to establish the timbral identity of sounds. As discussed in Subsections 3.3.1 and 3.3.3, the processes of timbral recognition may be analogous to a system of measurement space, through a feature space to an high level classificatory space. The spectral measurement space detailed in Chapter 4 produces partial tracks from which a feature space representation is extracted, as detailed in this chapter. Furthermore, this chapter elaborates on how the extracted features may be examined using statistical methods, for purposes of finding which of those in the extracted set are most effective in distinguishing between particular sound types.

It is not possible to mimic the entire process of timbral feature extraction in the human hearing system at the present state of knowledge. This research only considers a subset of the potentially huge range of features which could be investigated. As such, as with other aspects of this research, the consideration focuses on a limited timbre space. It does, however, consider a considerable number of features compared to previous studies in this area (Section 2.8.3).

There is a large amount of cortical processing between the acoustical and perceptual forms of sounds. This means that, while almost any measure of the spectral form will indicate something physical occurring, it does not necessarily follow that a perceptible change in

timbral quality will relate directly to it ([67]). Also, equivalent physical changes in different parts of the acoustical form are not certain to lead to perceptually equivalent magnitudes of effect. Spectral features are developed by considering those previously found to be perceptually important (Sections 2.10 and 2.11) and expanding the scope to encompass related aspects. This chapter confirms previous research and also expands knowledge of the links between perceptual and acoustic information.

Although the universality of results developed in a limited timbre space is questionable, the results which are obtained indicate much about the sort of relationships which exist between the acoustical and perceptual forms. Also important are the methods which are developed which could be used in future research and other timbre spaces.

“...the acoustic cues that signal instrument identity vary across contexts, and no one cue has been found that is necessary and sufficient for accurate identification.” [154]

It has been established in other chapters that:

1. The timbre of sounds is perceived in a logical, structured manner, with seemingly continuous discrimination between different sound qualities (Chapter 3).
2. Perception of timbre is an inherent part of the human experience and so common cortical processes must be at work (Chapter 3).
3. Previous researchers have found evidence of links between acoustical features and timbral attributes (Chapter 2).

These points mean that it is a viable proposition to attempt to find the features which are used by all humans to understand sound timbre. The essential problem is that, although certain areas of the spectrum have been established to be important (Section 2.10), there are a large number of other aspects which may also be important, but which have not yet been considered by researchers. However, it is known that the spectrum displays redundancy through correlation of different elements ([80] and Subsection 2.7.1). That is, it is not necessary to consider every possible parameter which may be derived from the spectral form in order to establish its timbral identity.

The feature extraction and analysis is designed to establish a number of things within the context of this research:

1. To show that features can be extracted from the spectral form which enable sound qualities to be distinguished; and that those features are not arbitrarily linked to perceptual differences, but show logical relationships. Although this has been shown in part in previous research (Subsection 2.8.3), it has not been tackled on a scale as large as this before, with so many stimuli and parameters.
2. To show that hierarchical discrimination of sound qualities is an effective system for structured decomposition of timbre space. As discussed in Subsection 2.7.2, there are different ways of constructing a timbre space representation for purposes of analysis and manipulation. This necessarily also indicates different ways of modelling the structure of timbral information in the auditory cortex. Whether an hierarchy of embedded distinctions is preferable to describing timbre space with a plain multidimensional axis set is of particular interest for future investigations.
3. To show that the dimensionality of timbre space is considerable and that the perceived differences between all perceptible stimuli cannot be completely explained by a low number of dimensions (3-5). Because this research deals with a wide range of sound qualities as well as more minor differences, it allows the consideration of how correct some authors have been in assuming that dimensionality results obtained from a small set of stimuli can be effectively transposed to the consideration of all sound (Subsection 2.7.1).

5.3 Major Elements of Feature Extraction

“The correlation between human perception and the related acoustic parameters is complex and not clearly understood.” [146]

This chapter is concerned with the characterisation of the spectral form in order to examine the parameters which may be of importance in distinguishing between sound stimuli based on their timbres. Specific values of features are measured in the spectral form of the stimuli, which are used as distinguishing axes. This is fundamentally different

from showing that a general area of the spectral form has some overall degree of difference between stimuli. For example, the statistical method of correlation can describe the similarity of two time-varying parameters without saying exactly what it is that is different. The amplitude envelope of stimuli can be a complex curve; showing that there are differences between stimuli's envelope using correlation is straightforward, but does not describe which parts of the envelope are different, nor quantify the differences in those parts relative to other stimuli in the data set.

Correlation techniques have been used previously, for example by Langmead ([107]). This research, however, attempts to establish particular features which characterise the more general areas. For example, the length of the attack portion is a particular feature of the amplitude envelope. Such an approach has been used elsewhere, in studies such as [98]. The advantage of measuring static values of specific parts of the spectral form must however be balanced against the attendant loss of information from the other parts, and the problems of characterisation of complex forms. As mentioned in Section 2.10, it is hard to quantify aspects of the spectral form which can only be described in terms such as "attack quality", or "random aspects", or "inequalities of acoustic form". It is apparent what the authors are trying to say, but only specific features can be objectively examined by mathematical methods. The task is to develop a set of specific parameters which cover the aspects of interest (Section 2.9). This is aided by the information in Sections 2.10 and 2.11, but the aim is also to attempt to broaden knowledge by consideration of related parameters.

"[There is a] . . . need for timbre perception models that reflect both static and time-varying properties of sound." [107]

There is an additional problem to feature extraction in that quantities are often time-varying. In the case of the overall amplitude, the properties of the envelope have been examined by other researchers. That is, features can be consistently located in those curves which are known to be perceptually important. However, there are many more quantities which can be extracted from a single frame of the spectrum, and so vary over time, which have no such known features. It is thus necessary to consider those in more general terms, as described by statistical measures such as the arithmetic mean, standard deviation and so on. Parameters can therefore be classed as "static" (denoted by S), which

correspond to those features which can be described by a single value, and those that are “time-varying” (\mathcal{T}), which must be described by the series of statistical measures for purposes of analysis detailed later in the chapter.

As mentioned in Section 5.2, this part of the research is also concerned with hierarchical decomposition. That is, the study is not concerned with trying to find a collection of features which can simultaneously differentiate between any stimulus and any other in a perceptually logical manner, but rather sets of features which progressively focus on areas of perception. This is not only an interesting structural form, but also aids in understanding the complex interacting elements of the spectrum and how they relate to perceived groupings. The long-term goal of research such as this is a timbre space described by SCAs (Section 2.4), where the SCAs are refined combinations of features, and the features are like those considered in this chapter. In that manner, SCAs would correspond to the major axes of difference within the timbral form, for purposes of gaining control of timbre space from a perceptual, rather than a spectral perspective.

All extraction of features is achieved algorithmically, rather than using any human intervention. This necessarily places a greater burden on the method to achieve the required results, but has the advantage of being objective. The methods used are, therefore, repeatable and should be directly comparable with other techniques developed in future. Additionally, if procedures have no additional intervention and are found to be consistently important in distinguishing particular sound qualities, then they may have a strong relationship to the actual processes at work in the cortex.

The extracted features are not intended to be an orthogonal set. That is, the vectors in timbre space that they represent are not independent measures but have overlapping effects. They have been chosen as they represent important viewpoints on the data being characterised. Similarly, they are not expected to represent an optimal final solution. Due to limitations of time, this research covers only some areas of the spectral form. This necessarily restricts the timbre space under consideration. The following areas are covered:

1. Static measures of amplitude envelope form. These concern specific measurements of the attack portion, which is known to have particular importance; but also the general form of the envelope, such as the number of large peaks and troughs.

2. Time-varying measures of strong partial characteristics. Strong partials are those with low bandwidth and high amplitude. Of particular interest are harmonics, but strong inharmonic elements are as well. Features concern such forms as the proportion of harmonic partials and frequency spread of strong harmonics.
3. Time-varying measures of spectral shape. That is, the general characteristics of the spectral envelope, such as the balance of amplitude between different spectral regions, and the spectral centroid.

Each of the areas above could be considered in a vast number of different ways. Those that have been chosen (Section 5.4) are designed to cover a considerable range of possible effects within the spectral form. They develop some of the themes considered in Sections 2.9, 2.10 and 2.11. The notable omissions from specific consideration are as follows:

1. Formant structures. These are particularly important in the study of vocal sounds, which are not covered by the stimuli used in this research. Establishing formants from a single instance of many other source types is particularly difficult to achieve with great accuracy.
2. Sustain/release portions. As described in Section 2.9, achieving a consistent evaluation of which parts of the stimuli are the sustain and release portions is very hard. Whereas, the attack portion can be more effectively established in most sounds. An alternative may be to devise a metric for the existence of sustain and only analyse for sustain/release properties in the appropriate stimuli. However, such considerations could get very complex, depending upon the relative importance of the decaying section of a sound with or without a sustain, and so on. Rather, the offset is treated with appropriate weighting in determining perceived timbre if everything after the attack is considered as one element, as in this research.
3. Spectral element widths, noise bands. This research does not consider how bandwidths of elements within the spectrum relate to timbral form, but does consider more general measures of spectral width, such as the spread of strong partials, and “noise”, such as the proportion of partials which are harmonic.
4. Synchrony of partials. The area of synchrony has been found to have some importance in previous research (see Section 2.9). As with formants, this is also a

complex area to achieve consistent objective analysis, and beyond the scope of this investigation.

5.4 Details of Feature Extraction



The data forms which represent timbral features within the time-varying frequency spectrum are only partly understood. With such a complex structure, perfect answers to the problems of characterisation do not exist. As such, the feature extraction detailed here is a particular set of methods using some elements which have been used before, plus an amount of new ideas, all modified through the process of experimental development. There are few actual algorithms detailed in the available literature. The feature extraction techniques given here represent forms which were developed by the author over a period of time to improve their characteristics relative to a range of testing stimuli from simple tones to complex instrumental sounds.

It is apparent from investigating the processes of feature extraction that there are a considerable number of ways of approaching every such problem. On the one hand, the aim is to achieve a simple interpretable set of axes, such that the results of searching for distinguishing acoustical aspects between perceptual groups can be understood. On the other, there is a requirement for the most timbrally appropriate features to be used (which can be very complex), or the results might be meaningless. The solution is to have a considerable number and range of axes, and tools which can find the combinations of interpretable features which are appropriate. In addition to this problem, there are the difficulties of defining acoustical measures for such perceptual attributes as the “complexity” within sounds ([149]). Also, it is necessary to produce feature extraction systems which are robust to the range of noise situations which present themselves in the acoustical forms of the stimuli. All these problems have been tackled in the features presented below.

An important aspect of the timbral feature extraction process is to decouple the consideration of timbre from the other elements of sound. In particular, amplitude, pitch and time/duration could cause the features to depend upon components other than timbral form, confusing the interpretation of results. To achieve decoupling, the features are often based on relative measures, such as proportions of the whole or ratios of values,

although some values necessarily need to be absolute measures. The process of developing features is essentially one of taking appropriate concepts and finding a way to generalise those metrics to make them comparable in analysis over the whole of the input set, without being forced into unnatural equalisation of the stimuli.

Other features in addition to those considered below were extracted in the process of investigation. These were concerned with testing the algorithmic processes used, such as the number of low-level peaks removed from consideration in obtaining the amplitude envelope. These represent typical spectral parameters which are not directly related to timbre. Later in the investigation it was shown, as hoped, that such metrics were not useful in explaining differences between stimuli. Although this is only a minor examination of an intuitively obvious concept, it shows that it is not correct to extract arbitrary spectral forms and expect some structured timbral link.

The features described in this section are designed to develop previous knowledge (Sections 2.10 and 2.11), while attempting to broaden the consideration to new areas of investigation. Algorithms are displayed in single boxes  and code fragments in double boxes .

5.4.1 Fundamental Frequency

“Hundreds of pitch-period estimation schemes have been proposed and tested”
[69]

The fundamental frequency (f_0) is a very important component in spectral feature extraction, as it provides a reference position for a number of the most important features of the timbral form as it is presently known. It provides a location point such that different sounds can be compared taking account of relative frequency positioning. It is important to note that f_0 is not always the same as the perceived pitch of the sound. In particular, pitch is the overall sensation of major oscillation frequency in a generally statically-pitched sound (such as those used in this research; Section 2.4). f_0 on the other hand is the major low oscillation mode of the acoustical form of the sound, which is also the lowest major common denominator of the harmonic form, but which can vary through even a statically pitched signal. This is confused by the following facts:

1. f_0 is not always physically present in the spectral form of a signal, but it may often be considered to exist through the pitch of the sound being close to its position, and/or the harmonic structure indicating its position.
2. Harmonics are not always at perfect integer multiples of the fundamental ([137]), yet are still perceived to be related to it.
3. The f_0 value may change through the signal and be present at some points (often the end of instrumental sounds when oscillation is usually stable) but not others (such as the start of those sounds).

Not all sounds have a fundamental frequency throughout the majority of the signal, but most of the stimuli in this research do. That doesn't necessarily make it an easy matter to find the f_0 values. It is apparent from studies of pitch and f_0 determination ([84], [4]) that a considerable number of such schemes has arisen from a lack of a simple and effective method which works in all situations, including providing consistent performance in areas of uncertain f_0 .

Restrictions on the timbre space described by a particular stimulus set can aid in improving the performance of feature extraction algorithms. For example, if the input sounds are only static portions of human male spoken vowel sounds, then large assumptions concerning the presence and amplitude of f_0 and harmonics and the likely range of f_0 can be made. Such assumptions aid greatly in the use of some systems such as many of those detailed in [84]. For a wider range of sounds, it is necessary to consider an approach which combines several metrics to achieve a reasonable result in a wide range of circumstances, which is what is used here. Basic f_0 finding techniques are generally either time-domain or frequency spectrum-based, although some hybrids exist ([84]). The technique in this research works in three stages as follows:

5.4.1.1 Stage 1 : Time-Domain Technique

This stage uses a method which attempts to match patterns of zeros and peaks in the input sound to find the fundamental period of oscillation at short steps along the sound. This has similarities to elements of techniques described in [69] and [33]. The basic algorithm is as follows:

- While not reached the end of the time-domain file
 - Search for ZCROSSINGS zero crossings and PEAKS peak positions from present position
 - If find zero crossing, store gradient and position
 - If have found the first zero, and find peak position, store position and amplitude
 - While not at end of file and not found a match for zero pattern
 - Search for next (test) set of ZCROSSINGS
- 1: - Sum differences between first and test set (for differences between each zero position and the first zero) = zposdiff
- 2: - Sum differences between first and test set zero gradients normalised by first set gradient = zgraddiff
 - Match found if zposdiff<ZPDIFFLIMIT and zgraddiff<ZGDIFFLIMIT (period value is difference between test and first positions)
 - Start at first set position again
 - While not at end of file and not found a match for peak pattern
 - Search for (test) set of PEAKS following the next zero crossing
- 3: - Sum differences between first and test set (ratios between peak to first zero position differences : first peak to zero difference) = pposdiff
- 4: - Sum differences between first and test set (ratios between peak amplitudes and first peak) = pampdiff
 - Match found if pposdiff<PPDIFFLIMIT and pampdiff<PADIFFLIMIT (period value again difference between test and first positions)
 - Vote output with shortest length
 - Move to next position

Code fragments:

```

1: zposdiff:=0.0;
   for arrpos:=2 to ZCROSSINGS do
     zposdiff:=zposdiff+Abs(
       (zeroarray[arrpos].samplenum-zeroarray[1].samplenum) -
       (testzeroarray[arrpos].samplenum-testzeroarray[1].samplenum));

```

```

2: zgraddiff:=0.0;
   for arrpos:=1 to ZCROSSINGS do
       zgraddiff:=zgraddiff+Abs(
           (zeroarray[arrpos].gradient - testzeroarray[arrpos].gradient) /
           zeroarray[arrpos].gradient);

```

```

3: pposdiff:=0.0;
   fsttestdiff:=testpeakarray[1].samplenum-testzeroarray[1].samplenum;
   fstorigdiff:=peakarray[1].samplenum-zeroarray[1].samplenum;
   for arrpos:=2 to PEAKS do begin
       origratio:=
           ((peakarray[arrpos].samplenum - zeroarray[1].samplenum)/
           fstorigdiff);
       newratio:=
           ((testpeakarray[arrpos].samplenum-testzeroarray[1].samplenum)/
           fsttestdiff);
       pposdiff:=pposdiff+Abs(origratio-newratio);
   end;

```

```

4: pampdiff:=0.0;
   for arrpos:=2 to PEAKS do begin
       origratio:=
           (peakarray[arrpos].amplitude/peakarray[1].amplitude);
       newratio:=
           (testpeakarray[arrpos].amplitude/testpeakarray[1].amplitude);
       pampdiff:=pampdiff+Abs(origratio-newratio);
   end;

```

The algorithm combines several ways of looking at the landmark points within the data set, thus the individual metrics are quite complicated. This combination required a considerable amount of development. As with the stability metrics of Chapter 4 it was found that a single metric was insufficient to produce reliable results. The balanced constants are as follows:

ZCROSSINGS = 6; PEAKS = 4; ZPDIFFLIMIT = 5.5; ZGDIFFLIMIT = 5.5;

PPDIFFLIMIT = 3.0; PADIFFLIMIT = 2.0;

5.4.1.2 Stage 2 : Frequency-Domain Technique

Having a spectral form available from Chapter 4 makes finding f_0 in the frequency domain easier than having to start from scratch. Because the aim is to find an oscillating pattern, the technique must find strong integer-related partials (harmonics) to indicate the position of f_0 . This is a so-called comb or frequency-template technique. This has similarities to elements of techniques described in [4] and [84].

The algorithm uses a set of comb weights based on the function from [84]:

$$\begin{aligned} \text{combweight}[\text{harmonic}] &= \text{harmonic}^{\frac{-1.0}{\text{COMBAMPFAC}}}, \\ \text{harmonic} &= 1 \dots \text{COMBSIZE}, \text{COMBAMPFAC} > 1 \end{aligned} \quad (5.1)$$

These are used to place more emphasis on the lower harmonics. The algorithm also reduces the partials under consideration to those which are considered “strong” in the context of this research. This means both harmonics and inharmonics that have low normalised bandwidth and moderately high amplitude (Hess recommends -35dB in [84]) relative to the maximum amplitude partial within the frame. The appropriate code fragment is:

```

for oscilcount:=1 to curoscount do begin
  if (initem^.values[AMP]<(biggestamp/AMPLIMITFAC)) or
    ((initem^.values[HBW]/initem^.values[FRQ])>NORMHBWCUTOFF)
  then begin
    (* remove item *)
    ...
  end;
end;

```

The basic algorithm is as follows:

- Generate comb weights
- For all spectral frames
 - Reduce spectral frame to strong partials
 - Scan for fundamental using comb method
- 1: - Scan through frequency range using (large) logarithmic steps
- 2: - Generate metric based on harmonic weight and proximity of partials to harmonic positions (with 1/x law roll-off).
- 3: - Scan through frequency range in region of first metric maximum in smaller steps for greater accuracy
 - Generate metric
 - Fundamental is metric maximum

Code fragments:

```
1: freqval:=LOWESTF;
   logfreqval:=Ln(freqval);
   freqvalmax:=halfsrates;
   while freqval<freqvalmax do begin
       Innercombscan(...);
       logfreqval:=logfreqval+LOGFRESOLUTIONL;
       freqval:=Exp(logfreqval);
   end;
```

```

2: totalres:=0.0;

for harmo:=1 to COMBSIZE do begin
    harmofreq:=freqval*harmo;
    initem:=inlisthead;

    for oscilcount:=1 to curoscils do begin
        (* use 1/x law normalised frequency metric *)
        totalres:=totalres+ ((combweight[harmo]*initem^.values[AMP])/
            (1.0+((Abs(initem^.values[FRQ]-harmofreq)/harmofreq)*
                CROLLOFFWEIGHT)));
        initem:=initem^.next;
    end;
end;

if totalres>greatestres then begin
    (* mark position if highest result thus far *)
    greatestres:=totalres;
    greatestf:=freqval;
end;

```

```

3: freqval:=Exp(Ln(greatestf)-LOGFRESOLUTIONL);
    freqvalmax:=Exp(Ln(greatestf)+LOGFRESOLUTIONL);
    highresstep:=(freqvalmax-freqval)/FRESOLUTIONHFAC;
    while freqval<freqvalmax do begin
        Innercombscan(...);
        freqval:=freqval+highresstep;
    end;

```

The balanced constants are as follows:

```

AMPLIMITFAC = 40; NORMHBWCUTOFF = 0.01; COMBSIZE = 10; COMBAMPFAC = 1.5;
CROLLOFFWEIGHT = 25.0; LOGFRESOLUTIONL = 0.02; FRESOLUTIONHFAC = 100;
LOWESTF = 20.0;

```

5.4.1.3 Stage 3 : Combination of Metrics

“...looking at zero crossing points or peaks alone is unlikely to produce a

robust pitch extraction technique

...

because of the effect of resonances on the frequency spectrum, pitch extraction techniques should not make assumptions about the presence of particular harmonics or sequences of harmonics" [4]

The techniques described in 5.4.1.1 and 5.4.1.2 were developed and tested independently on a number of sounds, but neither on their own proved as robust as desired. In particular the classical problem ([84]) in fundamental-finding techniques of octave errors was apparent, particularly in the comb (spectral) technique. A hybrid method is thus sensible. In addition to the zero/peak time metric and comb spectral metric, the frequency position of the maximum amplitude strong partial was also found to be a useful additional metric. A test-and-vote scheme is used; the test because values need a tolerance measure to achieve a match, and a vote rather than an average due to the sometimes large discrepancy between values. An typical test is as follows:

```
f2abottom:=f2ametvals.value/F2ARANGEFAC;
f2atop:=f2ametvals.value*F2ARANGEFAC;
if ((combscanval*matchfactor)>=f2abottom) and
  ((combscanval*matchfactor)<=f2atop) then begin
  Combscan:=combscanval*matchfactor;
  foundmatch:=true;
end;
```

In the example, the position of the zero/peak time metric (`f2ametvals.value`) extended by a tolerance factor is tested against the comb spectral metric, which can be changed by the integer octave multiplier/divisor `matchfactor`. The basic overall algorithm incorporating these tests is as follows:

- For all frames of data
 - Search for matches up to OCTMATCHLIMIT times the frequency of the metric being checked and similarly in the opposite direction
 - Test Stage 2 value multiplied by matching factor against Stage 1 value extended by tolerance factor F2RANGEFAC
 - Test Stage 2 value divided by matching factor against Stage 1 value (extended)
 - Test position of maximum amplitude * matching factor against Stage 1 values (extended)
 - Test position of maximum amplitude / matching factor against Stage 1 values (extended)
 - If previous tests fail
 - Test Stage 2 value (extended by F2RANGEFAC) against position of maximum amplitude
 - If all tests fail, take average of all the metrics
- Remove spurious values by checking for values of above (AVLIMITWIDTH times the region average), or below (region average divided by AVLIMITWIDTH) within regions of REGIONSIZE outputs
- Remove spurious values by replacing with the values most like others within a moving window of VWINSIZE outputs.

The final part is governed by a normalised $\frac{1}{x}$ relationship similar to that used in Stage 2:

```

for wincount:=1 to VWINSIZE do begin
  totalres:=0.0;
  for innercount:=1 to VWINSIZE do begin
    if (winvals[innercount]>0.0) and (innercount<>wincount) then
      totalres:=totalres+(1.0/
        (1.0+((Abs(winvals[innercount]-winvals[wincount])/
          winvals[wincount])*VROLLOFFWEIGHT)));
  end;
  if totalres>greatestres then begin
    greatestres:=totalres;
    voterresult:=winvals[wincount];
  end;
end;
end;

```

The balanced constants are as follows:

```

F2RANGEFAC = 1.4; OCTMATCHLIMIT = 2; REGIONSIZE=15; AVLIMITWIDTH=1.3;
VWINSIZE = 10; VROLLOFFWEIGHT = 1.5;

```

5.4.2 Time-Varying Measures of Strong Partial Characteristics

It is known that timbre is influenced by the harmonic partials of a stimulus (Section 2.9). The concept of strong partials (as derived in Subsubsection 5.4.1.2) extends that basic principle to encompass inharmonics with low bandwidth and moderate amplitude in the same group of features. This is because it is known that inharmonics also have an important role in determining timbral form.

The first stage of developing these features is to find the harmonics, based on the value of f_0 derived using the technique of Subsection 5.4.1. This is achieved by searching the strong partials for multiples of f_0 . However, as mentioned in Subsection 5.4.1, the harmonics of traditional instruments are not always perfect multiples of the fundamental, which can relate to such perceptual aspects as the warmth of the sound (Section 2.11). Matches are achieved, therefore, by scanning for partials with an increasing range of (multiplicative) tolerance factor in steps of TOLERANCESTEP up to MAXTOLERANCE (with values of 0.0025 and 0.05 in the final version).

The number of harmonics being considered in this research is the same as the COMBSIZE used in finding f_0 ; that is, 10. This value was chosen due to the known importance of the lower harmonics (Section 2.9) and how the higher harmonics (above the 7th) tend to be more grouped together in perception. Over a considerable range of stimuli, such as those used in this research, a large amount of noise and null data can creep into the consideration if research considers large numbers of harmonics in general ways as described here.

Extracted Feature 1 : Proportion of all partials that are harmonic (\mathcal{T})

This feature gives an indication of whether the stimulus tends toward a simple oscillatory form. The feature uses all the partials in the sound not just the strong ones. This is a very simple metric and is of use with simple sounds, or those regions of a stimulus where the oscillation has reduced from a complex to a steady form. It might be expected that this value would increase through a struck instrument sound, for example.

Extracted Feature 2 : Proportion of all partials that are strong inharmonics (\mathcal{T})

This has a similar type of role to the previous feature, but is not simply the opposite, in that it measures the number of strong inharmonics, not just anything that isn't harmonic. Such partials have an important role in timbre sensation.

Extracted Feature 3 : Proportion of strong partials that are harmonic (\mathcal{T})

This completes the set of three basic proportions. It is desirable to have a number of ways of looking at the spectral components in this manner, to find which particular ways of viewing them is important. This is demonstrated in the other choices of features as well. Interest lies not only in the fact that harmonics, say, are generally of interest, but relative to which other aspects, and how.

Extracted Feature 4 : Average inharmonicity (\mathcal{T})

The match to the strict integer-multiple harmonic template is regularly mentioned in the literature (Sections 2.10 and 2.11). The way the harmonics are stretched can affect components like “warmth” in the sound. In this research, the value of the feature corresponds to the average tolerance factor used to match the harmonics.

Extracted Feature 5 : Average inharmonicity (weighted) (\mathcal{T})

This is the same as the previous feature, but includes weightings to place more emphasis

on the lower harmonics. The same weightings are used as with the comb in 5.4.1.2.

Extracted Feature 6 : Strong inharmonic to harmonic proximity (\mathcal{T})

The proximity of partials leads to timbral effects, such as roughness when close. In this research the algorithm takes each strong partial and finds the minimum \log_{10} frequency difference to the nearest harmonic. The feature value is the sum of the $\frac{0.001}{0.001+m\text{indiff}}$ values, so larger values imply greater proximity.

Extracted Feature 7 : \log_{10} difference frequency spread of harmonics (\mathcal{T})

Extracted Feature 8 : \log_{10} difference frequency spread of strong partials (\mathcal{T})

Extracted Feature 9 : Upper frequency of strong partials (\mathcal{T})

Extracted Feature 10 : Lower frequency of strong partials (\mathcal{T})

These four features are indicators of spectral spread in the harmonics and strong partials. The two viewpoints are logarithmic and linear for comparison purposes; logarithmic possibly being more aurally “appropriate”, given the analysis shape of the peripheral hearing system (Chapter 4).

Extracted Feature 11 : Harmonic to fundamental relative amplitudes (9 features) (\mathcal{T})

Rather than pick out particular harmonic patterns to represent the balance of harmonics in the stimuli, this research extracts a feature for each of the relative amplitudes of the harmonics to the fundamental (the 1st harmonic). Through the techniques described in Section 5.6, the importance of different combinations of harmonics can become apparent.

5.4.3 Amplitude Envelope

The overall amplitude curve of stimuli is a very important source of timbral information. It is most accurately derived from the original time-domain form of the signal. The process requires the removal of higher frequency components to leave a smooth trajectory. However, merely applying a low-pass filter to the data results in an trade-off between transient performance in areas where rapid change is of importance to the ear, such as the attack, and removal of noise from smooth stages. As such, it is desirable to be more dynamic in the derivation technique.

In this research, f_0 is used as the indicator of appropriate length of region of consideration. This is naturally related to the stability/periodicity metrics of Chapter 4 and has a similar effect in indicating the dynamics of the form. $PERNUM/f_0$ is used as the length over which peaks in the time-domain form are searched for. A previous version of this derivation used a number of zero crossings instead, but that was not nearly as effective. The basic algorithm is as follows:

- Step through the input data in PERNUM period lengths (lengths of consideration)
- Search in the length of consideration for maximum peak in absolute values of input signal
- Ignore peaks less than the amplitude threshold (AMPTHRESH parts in 32767)
- Ignore peaks within +/- NOISETHRESH parts in 32767 of the previous peak value, to remove unnecessary duplicate data
- Remove peaks resulting in short distance gradient direction changes +ve/-ve/+ve or -ve/+ve/-ve.

The balanced constants are as follows:

$PERNUM = 10.0$; $AMPTHRESH = 20.0$; $NOISETHRESH = 5.0$;

5.4.4 Static Measures of Amplitude Envelope Form

From the smoothed amplitude envelope, a number of characteristics are derived in this research relating to the important aspects of its shape. As can be seen from Sections 2.10 and 2.11, previous research concerning the nature of the amplitude envelope has found the onset constantly important, and the general tendency of the envelope of interest, yet specific details have been a little scarce. Some of the features detailed here are quite new ways of viewing the nature of the envelope, compared to those referenced in 2.10.

Extracted Feature 12 : Attack time (S)

The actual value of attack time has been known to be important for some time. It is defined here as the time to the first major peak in the amplitude curve, rather than being the position of highest value within a particular length as has been used before (such as

the first 50ms, as in [107]).

Extracted Feature 13 : 10-90% proportion of attack (S)

The proportion of the attack time taken to move from 10% to 90% of the amplitude at the attack peak gives an indication of the shape of the attack. Linear attacks will result in value of 0.8, whereas slow starting attacks will produce lower values, and so on.

Extracted Feature 14 : Attack to average ratio (S)

The ratio of attack amplitude to average level indicates the general shape of the entire curve, for distinguishing those sounds which have a more percussive nature, for example.

Extracted Feature 15 : Largest value position to attack position ratio (S)

Where the attack peak is not the maximum level in the amplitude curve, this value will differ from 1. This is a speculative feature based on the observation of some waveshapes.

Extracted Feature 16 : Number peaks and troughs (S)

The number of peaks and troughs is one indicator of the “complexity” of the amplitude form. For example, a simple test tone will have no amplitude ornaments, whereas a sound with vibrato will have a number of major peaks and troughs, and a complex natural sound may have many variations.

Extracted Feature 17 : Breakpoints (S)

This feature is also designed to facilitate explanation of some of the complexity in sounds. In the envelope derivation, breakpoints are only placed where a significant change in level is observed, so simple envelopes will have fewer breakpoints than those with many inflections.

Extracted Feature 18 : Average absolute gradient (S)

This is the third complexity measure and relates to the “smoothness” of the envelope. Curves can have the same average gradient, yet very different absolute averages, due to the number of inflections. This is complementary to the previous two in that it also imparts amplitude variation information.

Extracted Feature 19 : Amplitude envelope proportions of low mode magnitudes in first 20 modes (20 features) (S)

Extracted Feature 20 : Amplitude envelope proportions of low frequency

magnitudes in first 20 interpolated frequencies at 1.5Hz spacing (20 features)
(S)

These two sets of features are a general description of the shape of the amplitude envelope from the point of view of modes of vibration (that is, length-independent oscillatory modes) and low-frequency amplitude values at 1.5Hz spacing. Both are derived from a DFT of the envelope data; the latter is based on linear interpolation of the DFT results. They are then converted to proportions within the each set of 20 to make them amplitude-independent.

5.4.5 Time-Varying Measures of Spectral Shape

The general shape of the spectral form is important in indicating such aspects as where the resonances exist.

Extracted Feature 21 : Slope metric (lin a, log₁₀f) (T)

Extracted Feature 22 : Slope metric (log₁₀a, log₁₀f) (T)

Measures of spectral slope have previously been found to be of importance. The two features extracted in this research are based on a linear regression (see Appendix B) through the points corresponding to the strong partials in the amplitude spectrum. The strong partials are extracted using the same technique as described in 5.4.1.2 and the result metric is the gradient of the regressed line. Two metrics are produced, one based on linear amplitudes and the other on log₁₀ values. Both metrics use log₁₀ values of frequency.

Extracted Feature 23 : log₁₀ centroid (T)

Extracted Feature 24 : log₁₀ power centroid (T)

The spectral centroid is another common spectral feature in the literature. The metric calculation is based on all partials rather than just the strong ones used with the slope metrics. Centroids are calculated using the following equations with linear frequency:

$$\text{centroid} = \frac{1}{\sum_{x=1}^N a[x]} \sum_{x=1}^N a[x]f[x] \quad (5.2)$$

$$\text{power centroid} = \frac{1}{\sum_{x=1}^N a^2[x]} \sum_{x=1}^N a^2[x]f[x] \quad (5.3)$$

Extracted Feature 25 : Split bin amplitude proportions (8 features) (T)

The split bin amplitude proportions features are a simple gross overview of the distribution of emphasis in the spectrum. They are based on the sum of amplitudes of all the partials in 8 logarithmically-spaced bins. The upper frequency of the bins is calculated from:

$$upperbinedge[bin] = 10^{(bin * expfactor) + \log_{10}(20)}, \quad expfactor = \frac{\log_{10}(11025) - \log_{10}(20)}{8},$$

$bin = 1 \dots 8$ (5.4)

The features are the proportions of the total amplitude in each of the bins.

Extracted Feature 26 : Peak:average spectral amplitude ratio (\mathcal{T})

This measure is designed to facilitate examination of whether the spectrum tends to be flat, or has large features above the general amplitude level.

5.4.6 Development and Testing Procedure

The feature extraction algorithms detailed in this chapter were developed and validated with reference to testing material to investigate their performance with stimuli having particular characteristics. This was necessary to show that the features could be consistently and automatically extracted by the algorithms from the analysed forms which were generated with the analysis-synthesis tool of Chapter 4. To facilitate this, each feature extraction algorithm has an associated graphical display program, to allow examination of the result data.

The data that should result from the feature extractions is apparent in the graphical plots of the time domain and spectral data. These show the strong data forms, which can then be compared with the algorithmically-extracted information. The stimuli used to test the algorithms display particular strong characteristics by design, but the features present in the full set of 153 input stimuli (Appendix A) can be located by eye in the graphical plots. As such, and also due to limitations of time, the author did not employ other subjects for assessment of these results. The different feature extraction stages were tested as follows:

1. Fundamental frequency (Subsection 5.4.1).

The pitch of the stimuli fed into the system was known in advance, which indicated the region of the fundamental frequency, as did the harmonic forms visible in the data. This vital part of the feature extraction process was developed by considering

different complexities of test stimulus from sine tone, to purely harmonic test tones, highly harmonic non-synthetic instrument tones, and finally more ambiguous tones such as drum sounds where large wideband frequency content existed. It was very important to get this part right due to the way fundamental frequency is used in the subsequent extraction processes. It thus received the greatest amount of development iterations.

2. **Time-varying measures of strong partial characteristics** (Subsection 5.4.2).

Isolating strong partials (low bandwidth and moderate amplitude spectral elements) from other parts of the spectral form permits examination of harmonic and inharmonic structures. To test this process, as with the fundamental frequency extraction, it was a case of using stimuli with increasing complexity from sine tone, to harmonic test tones, to harmonics and inharmonics, and then onto less synthetic tones. It was particularly important that these elements could be found effectively in the presence of other components.

3. **Amplitude envelope** (Subsection 5.4.3).

The amplitude envelope for test tones should be found to be a clean rectangular shape. With more dynamic profiles such as percussive sounds, the tester is looking for rapid performance in high-energy stages and smooth trajectories in relaxed stages. That is, it is necessary to compare with the shape of the unsmoothed time-domain input to check that the high-energy stages are not smoothed-out (reducing the peaks), and that the relaxed stages are smoothed enough to remove localised noise elements. However, there is also a balance to be struck with sounds that display vibrato effects that these are not removed from the profile.

4. **Static measures of amplitude envelope form** (Subsection 5.4.4).

These measures are easily graphically represented and can be checked against the plotted smoothed amplitude envelope and the original time-domain form. These extractions were tested with the principal shapes of envelope; rectangular forms, percussive forms, “classical” attack-decay-sustain-release forms, forms with vibrato, and more complex forms. The latter includes dynamic elements later in the sound than the attack, variation of noise content through the sound and so on.

5. **Time-varying measures of spectral shape** (Subsection 5.4.5).

It was again possible to test these feature extractions by the use of particular stimuli

with the appropriate characteristics. A few partials can be used to check that the algorithms find slopes, centroids and peak:average ratios successfully. The simplicity of these features means that the method scales very simply and can be shown to do so with more complex sets.

After testing the individual feature extraction sections with particular stimuli to assess performance, the feature extraction was performed on all 153 input sounds used in this research. The results were then assessed and deficiencies found in a small number of the features. Those algorithms were modified and the feature extraction repeated.

5.4.7 Feature Extraction Summary

This section has described the features extracted from the time-varying frequency spectrum forms of the stimuli used in this research, which are used in the following sections to analyse the nature of timbral difference between those input sounds. They are summarised in Table 5.1. They encompass a wide range of measures within the areas specified in Section 5.3, which in turn were specified based on an extrapolation of the data presented in previous research (Sections 2.9, 2.10 and 2.11).

| No. | Description | Type | Features |
|--|--|---------------|----------|
| Time-Varying Measures of Strong Partial Characteristics | | | |
| 1 | Proportion of all partials that are harmonic | \mathcal{T} | 1 |
| 2 | Proportion of all partials that are strong inharmonics | \mathcal{T} | 1 |
| 3 | Proportion of strong partials that are harmonic | \mathcal{T} | 1 |
| 4 | Average inharmonicity | \mathcal{T} | 1 |
| 5 | Average inharmonicity (weighted) | \mathcal{T} | 1 |
| 6 | Strong inharmonic to harmonic proximity | \mathcal{T} | 1 |
| 7 | \log_{10} difference frequency spread of harmonics | \mathcal{T} | 1 |
| 8 | \log_{10} difference frequency spread of strong partials | \mathcal{T} | 1 |
| 9 | Upper frequency of strong partials | \mathcal{T} | 1 |

Table 5.1: Summary of Extracted Feature Set

| No. | Description | Type | Features |
|-----|---|---------------|----------|
| 10 | Lower frequency of strong partials | \mathcal{T} | 1 |
| 11 | Harmonic to fundamental relative amplitudes | \mathcal{T} | 9 |

Static Measures of Amplitude Envelope Form

| | | | |
|----|--|---------------|----|
| 12 | Attack time | \mathcal{S} | 1 |
| 13 | 10-90% proportion of attack | \mathcal{S} | 1 |
| 14 | Attack to average ratio | \mathcal{S} | 1 |
| 15 | Largest value position to attack position ratio | \mathcal{S} | 1 |
| 16 | Number peaks and troughs | \mathcal{S} | 1 |
| 17 | Breakpoints | \mathcal{S} | 1 |
| 18 | Average absolute gradient | \mathcal{S} | 1 |
| 19 | Amplitude envelope proportions of low mode magnitudes in first 20 modes | \mathcal{S} | 20 |
| 20 | Amplitude envelope proportions of low frequency magnitudes in first 20 interpolated frequencies at 1.5Hz spacing | \mathcal{S} | 20 |

Time-Varying Measures of Spectral Shape

| | | | |
|----|--|---------------|---|
| 21 | Slope metric ($\ln a$, $\log_{10}f$) | \mathcal{T} | 1 |
| 22 | Slope metric ($\log_{10}a$, $\log_{10}f$) | \mathcal{T} | 1 |
| 23 | \log_{10} centroid | \mathcal{T} | 1 |
| 24 | \log_{10} power centroid | \mathcal{T} | 1 |
| 25 | Split bin amplitude proportions | \mathcal{T} | 8 |
| 26 | Peak:average spectral amplitude ratio | \mathcal{T} | 1 |

Table 5.1: Summary of Extracted Feature Set

5.5 Overview of Feature Set Analysis

“... multidimensionality creates difficulties for developing timbral intervals, because the factors relating one timbre to another are acoustically and perceptually so complicated.” [111]

Having extracted spectral features as described in Section 5.4, the next stage is to analyse those features' values for the 153 stimuli used in this research. The aim is to find out which features in the extracted set contribute most to the perceived differences between groups of stimuli. The groups are chosen to facilitate investigation of the links between perceived timbral structure and acoustical form.

The problem is to relate known perceptual differences to the spectral feature differences which exist. This means specifying which differences are perceived (that is, which stimuli are perceived as different or similar in that context) and searching for features which have systematic differences corresponding to that arrangement. Different searching methods can be applied to such a problem according to how the data points are believed to be related.

The overall plan is to find which features might be similar to those used by the auditory cortex in establishing timbral form. In reality, the timbre space developed in this research restricts such grand designs. Understanding the effectiveness of the methods tested here in finding those relationships is important, however, such that future research can improve on the techniques which are outlined. It is also of interest to find if the chosen features have the effects that might be expected, such as attack properties relating to percussiveness and so forth.

As with the perceptual study detailed in Chapter 3, a number of methods are employed in this chapter to establish an understanding of the data set from different perspectives. The statistical methods employed in this chapter are described and compared in Appendix B. The statistical algorithm developed by the author specifically for this timbral feature set analysis is described in detail in that Appendix and a brief overview is given in Subsection 5.5.3. Appendix B also lists references to papers which have used particular multivariate statistical techniques in investigating timbre. More specifically to the general style employed here, finding axes of timbral significance based on the statistics of sound sources has been demonstrated previously in such papers as [46].

5.5.1 Characterisation of Time-Varying Forms

As mentioned in Section 5.3, variation with time is present in a number of features which cannot be easily categorised with static metrics. In this research, single-value characteristic measures are extracted from all time-varying parameters, using general statistical

methods. This goes some way to isolating which part of the time-variation is of interest, rather than treating that variation as a complete unit by, say, correlating the features together, as was used in [107]. It is necessary as there is currently a lack of information concerning which parts of which time-varying forms are of interest. Note that these metrics are still different from the static ones, as they are general categorisations, not individually devised in the manner of the attack to average ratio metric.

The time-varying parameters are characterised by three viewpoints; as a whole, the onset part, and the release (non-attack part). Each of the three viewpoints is analysed by three methods; the arithmetic mean, standard deviation, and the slope/gradient of a line regressed through the data. These methods provide information on the average effects, the quantity of variation from those averages, and the overall tendency of the variation, respectively. As with extracting the spectral parameters from the available model data, there are an almost endless number of ways of characterising a data set. The techniques used here have been chosen because:

1. They are well known and can be compared easily in future studies.
2. They represent broad tendencies within the data which indicate where future investigations should concentrate.
3. They generate a small number of output values, which minimises the amount of computation necessary when using these metrics in feature analysis.

The time-varying metric types are coded in this chapter as shown in Table 5.2. In addition, the code SF is used to indicate a static feature, for clarity. When features are referred-to in the following text, they are given in the form FF.SSCC, where FF is the extracted feature number, SS is the subfeature where several parts exist in the same extracted feature, and CC is the code type. For example, 12.0SF is the (static) attack time, 20.2SF is the (static) 3Hz low frequency magnitude of the amplitude envelope metric, and 11.3SR is the standard deviation in the release part of the 3rd harmonic-fundamental relative amplitude.

| | Arith.mean average | Standard deviation | Regression gradient |
|----------|--------------------|--------------------|---------------------|
| All data | AA | SA | RA |
| Onset | AO | SO | RO |
| Release | AR | SR | RR |

Table 5.2: Time-Varying Characteristic Metric Codes

5.5.2 Dimensional Considerations

The result of generating each of the 9 types of statistical metrics for each of the time-varying features, plus the static features, is a total of 335 features which form the data set, with values for each of the 153 sounds of Appendix A. This is considerably more feature data than has ever been analysed at one time in a study of sound timbre. In comparison, the recent study by Langmead ([107]) considered 9 features (times 2 measurement techniques) for 16 sounds. While a data set with large scope leads to the possibility of conclusions and the development of methods which are more widely applicable, it also leads to more complex data relationships being present. The structure of relationships between a small number of perceptually distant stimuli can be successfully explained by a small number of well chosen parameters. A large number of stimuli displaying more similarities among them are not as well explained by the same group of distinguishing parameters; the structured form at an high level no longer explains the low-level discrimination in a perceptually ordered manner. A larger/different set of features is required to achieve that lower-level discrimination, as shown in Subsection 5.6.2.

Additionally, the more features there are that compose the data set, the more important it is that the mechanisms for investigating the data space facilitate clear interpretation. For example, if the discriminant analysis technique (see Appendix B) is used to find contributions to grouping in a set of 335 features, then the result will be a set of 335 weighting values, which is very difficult to interpret.

The number of dimensions of change involved in the perception of timbre was discussed from a conceptual point of view in Subsection 2.7.1. However, it is necessary to know what dimensionality means in practical terms before examining the relationships in the data set. Assuming an unquantised and unlimited form, a single feature from spectral analyses can distinguish between the stimuli being considered in absolute terms. However, the

dimensionality of timbre space is determined by an ability to achieve an ordered structure of stimulus points which relates to perceptual difference. The dimensionality, or degrees of freedom, is determined by the best set of axes, in some sense, for distinguishing between groups of stimuli at all levels of detail (of interest) in a perceptually structured manner. In this research, a *subset* of axes are objectively chosen by an algorithm from the raw features (see Subsection 5.5.3). In some other research, the axes have been a scaled set of components derived from the input feature set using such techniques as Multidimensional Scaling and Principal Components Analysis (see Appendix B). The sense in which the dimensions found are best is determined by the perceptual accuracy of the distinctions between stimuli which are produced by the objectively determined feature subset, which in turn is determined by the features used as input and the algorithm which determines the dimensions that most appropriately achieve the structure.

5.5.3 Feature Set Searching Technique

The major algorithm used for searching the set of features for the most appropriate axes to describe differences between groups of input stimuli is coded in a program written by the author called AXESDIST. The algorithm is described in Section B.3 in the context of the other statistical methods of Appendix B. This subsection overviews the features of AXESDIST and its relationship to other techniques.

As described in Subsection 5.5.2, a tool is required to find the optimal set of axes to describe the relationship between perceptual structure and acoustical form. Other researchers (such as [71] and others referenced in Appendix B) have considered the structure of timbre space by scaling a perceptual structure into a low dimensional space and then comparing the resulting axes with acoustical dimensions. This concept is continued in Subsection 5.6.1. However, to investigate more complex relationships which cannot be scaled from the results of the perceptual study detailed in this thesis (Chapter 3) requires more direct examination of the acoustical forms which are associated with the perceptual grouping of stimuli.

By specifying groups of stimuli and the criteria by which they are to be separated, AXESDIST systematically searches the feature space for sets of axes which best achieve the desired containment within, and separation between, groups. AXESDIST performs an

exhaustive enumeration to find the features which minimise the following fraction:

$$\frac{\begin{aligned} & Within(nGroup_1, nMethod_1) + \dots \\ & + Within(nGroup_p, nMethod_p) \end{aligned}}{\begin{aligned} & Between(dGroup_{1a} \leftrightarrow dGroup_{1b}, dMethod_1) + \dots \\ & + Between(dGroup_{pa} \leftrightarrow dGroup_{pb}, dMethod_p) \end{aligned}} \quad (5.5)$$

The user specifies the *Methods* by which distance metrics are measured for containment (*Within*) and discrimination (*Between*) the groups of stimuli. The *nMethods* for containment are:

1. Average all points distance (averageall).
2. Maximum inter-point distance (maxdistall).

The *dMethods* for distinction are:

1. Average all points distance (averageall).
2. Minimum inter-point distance (mindistall).
3. Centroid distance (centroiddist).

AXESDIST finds a prespecified number of features as the solution from the original set of 335, rather than a complex scaled or weighted solution (where the answer is a mix of a number of features, rather than the features themselves) as some other techniques used in this area do. What this means is that the result is more interpretable in terms of the acoustical form. Overall, therefore, AXESDIST is a tool to aid in the examination of the relationship between the perceptual and acoustical forms from a user-specified perspective. Having specified the perspective, the algorithm takes no further user input to drive the data examination. It can thus be considered an objective technique. However, the result is necessarily a particular viewpoint on the data, not *the* answer.

The verification of AXESDIST's correct operation was achieved by stressing the tool with particular stimulus groups, plotting the feature results generated, and examining the distribution of data points compared to that requested. By restricting the feature set and stimulus set under investigation, and examining the arrangement of data points along

those features by hand, it is possible to show whether AXESDIST is indeed finding the appropriate features for most effective discrimination/containment. Stimuli can be chosen which display differences in a restricted subset of features, and AXESDIST was found to choose those features consistently. Also, as will be shown in the investigation of dimensional/noise effects later (Subsection 5.6.3), varying the feature set used in the algorithm restricts the available choices, and so the potential pertinence of the results from AXESDIST. Thus, as the feature set includes more and more relevant axes to the effect being considered, the discrimination/containment will improve if the algorithm is working correctly.

5.6 Details of Feature Set Analysis

The analysis of the feature space detailed here has a number of objectives:

1. Linking the structure of timbre perception to acoustical form.
2. Establishing that the dimensionality of timbre perception is considerable (i.e. must be described by more than 3-5 dimensions).
3. Showing that hierarchical decomposition is a valid and useful technique for exploring the acoustical/perceptual structure relating to timbre.
4. Understanding more about which acoustical components are linked to differences between particular sound types.
5. Developing methods for discriminating between sound groups.
6. Understanding the problems of examining the features of the complex data set.

A preliminary study conducted by the author showed that the individual features in the set of 335 were very hard to interpret by hand, which shows the importance of objective and effective computer-based pattern analysis techniques in this area. The researcher is then left with the problem of understanding the resulting computer-generated best fit solutions. As with the perceptual data in Chapter 3, the acoustical data does not normally demonstrate clean separated groups of data. In fact, that is the likely reason why perceptual confusions occur with limited stimulus data. Secondly, it is worth remembering

that the resultant sets of features from programs such as AXESDIST are usually a combinatorial effect; that is, it is the features acting together that achieves the distinction between timbre types. It is important, then, to be careful not be drawn into vast sweeping statements about the nature of the perception/acoustics of timbre, as the results can be harder to interpret correctly than might be initially assumed.

5.6.1 Correlation of Scaled Perceptual Dimensions with Acoustical Features

In Subsection 3.5.2, correlation was used to assess the direction, form and strength of the relationship between the different perceptual test results that were produced by the participants. This was a very informative exercise, particularly as it was relating very similar variables which were expected to have direct relationships. The same method (Pearson correlation, see Appendix B) can be used to assess whether there is a relationship between the principal component result dimensions of the PCA procedures of Subsections 3.5.4 and 3.5.5, and the 335 acoustical features described in Section 5.4.

This is, in fact asking a very different statistical question from that in Subsection 3.5.2. That use of correlation was to show that the results were related between and within participants' set of 6 tests. In this subsection, however, the use of correlation asks if single feature dimensions have the same form as single PCA solution dimensions. If there exists a small group of strong correlations with each of the PCA result dimensions, it could be said that the timbral differences among the 153 stimuli as represented by 2 or 3 dimensions can be explained successfully with a small set of simple acoustical features.

Previous authors (such as Grey, [71]) have estimated that this can be achieved. What stands in the way of this sort of result is that it assumes that timbre, as represented by the 153 sounds used in this research, can be explained by 2-3 features (one per dimension of the PCA solution) drawn from the 335 simple features used here. The alternatives are:

1. The 335 features do not encompass enough information about the acoustical form to represent the differences (which is dispelled by the results of Subsection 5.6.2 later).
2. There is too much noise in the derivation of the features to explain the differences (again, see Subsection 5.6.2).

3. That the PCA results are not rotated appropriately for comparison with single feature dimensions.
4. That timbre cannot be explained with such a simple dimensional form; that the PCA dimensions represent a complex perceptual structure which requires multiple simple features to effectively explain the differences in each PCA dimension.

The last of these seems most likely in the context of the other experiments detailed later in this chapter.

The highest correlation values between the 2D PCA results and the feature axes are found in the first factor dimension of the solution. These correlations are with features 19 and 20 (low mode and low frequency descriptions of the amplitude envelope). In particular, mode 0 (feature 19.0) and 0Hz (feature 20.0) usually display the highest correlation values (see Table 5.3). This is consistent with the expected distinctions between the Woodwind/Brass and Percussive/Hammered pairs displayed in Figure 3.9. This also relates to the high values correlations with feature 14 (attack to average ratio in amplitude envelope) in Table 5.3.

Higher mode/frequency correlation values for features 19/20 are also often above 0.5. Other features do not produce such high correlation values as 14, 19 and 20. There are some moderate correlation values which are also of interest, however. The majority of the consistently moderately correlated features for the first factor are given in Table 5.4. These are 3.0SO (proportion of strong partials that are harmonic, stddev/onset), 4.0SA (average inharmonicity, stddev/all), 5.0AA (average inharmonicity, weighted, average/all), 8.0AO (\log_{10} frequency spread of strong partials, average/onset), 9.0SO (upper frequency of strong partials, stddev/onset), 23.0SO (\log_{10} centroid, stddev/onset), and 24.0SO (\log_{10} power centroid, stddev/onset). This is a considerable range of measures and perspectives on those measures. However, they represent “classical” timbral dimensions which have been found before to be important general dimensions. For example, Grey’s interpretation of his results ([71]) produced dimensions of spectral energy distribution; low amplitude, high frequency energy in the attack (possibly inharmonic); and synchrony of attacks and decays of higher harmonics (or possibly musical instrument family relationships). These relate quite well to the sort of correlations being seen here.

Although some interpretable, strong axes are apparent in the correlations for the first PCA

| Participant | Feature 14.OSF | Feature 19.OSF | Feature 20.OSF |
|-------------|----------------|----------------|----------------|
| 1 | -0.587 | 0.701 | 0.742 |
| 2 | -0.666 | 0.804 | 0.747 |
| 3 | -0.443 | 0.552 | 0.602 |
| 4 | -0.536 | 0.629 | 0.672 |
| 5 | 0.597 | -0.681 | -0.778 |
| 6 | 0.418 | -0.535 | -0.558 |
| 7 | -0.579 | 0.731 | 0.751 |
| 8 | -0.561 | 0.631 | 0.713 |
| 9 | -0.607 | 0.691 | 0.719 |
| 10 | 0.585 | -0.704 | -0.765 |
| 11 | -0.638 | 0.832 | 0.762 |
| 12 | -0.592 | 0.714 | 0.757 |
| 13 | 0.582 | -0.683 | -0.748 |
| 14 | 0.601 | -0.711 | -0.716 |

Table 5.3: Perceptual Test 2D PCA (First Factor) versus Features: Highlights of Strongest Correlations

dimension, the second result factor displays no such consistency and few values of magnitude above 0.4. The author interprets this result as showing that the PCA has resulted in one factor which has the strongest dimensional form which can be correlated with the classical timbral features. The other factor is everything else; all the nuances of timbre in addition to the major dimension(s).

A similar pattern to the 2D case emerges with the 3D correlations. Again, there is a single axis of stronger correlations with the axes identified in the 2D case. The other two dimensions display a lack of strong correlations. If the nature of timbre were really a simple 2 or 3 dimensional structure explainable by the sort of features detailed here (which are derived from the results of previous researchers, some of whom postulated that concept) then it would be expected that the results of this section would have been a set of 2 or 3 fundamental axes of timbre. The author believes that the feature set in this research covers enough of the timbre space of interest and that the PCA results are best aligned for this to have happened, were it really the case. The reason that previous researchers have

| Participant | 3.0SO | 4.0SA | 5.0AA | 8.0AO | 9.0SO | 23.0SO | 24.0SO |
|-------------|--------|--------|--------|--------|--------|--------|--------|
| 1 | 0.387 | -0.370 | -0.404 | 0.547 | 0.436 | -0.413 | -0.419 |
| 2 | 0.285 | -0.480 | -0.392 | 0.415 | 0.304 | -0.352 | -0.359 |
| 3 | 0.381 | -0.324 | -0.388 | 0.482 | 0.351 | -0.343 | -0.343 |
| 4 | 0.414 | -0.317 | -0.351 | 0.520 | 0.448 | -0.426 | -0.427 |
| 5 | -0.423 | 0.385 | 0.439 | -0.577 | -0.441 | 0.488 | 0.492 |
| 6 | -0.272 | 0.286 | 0.356 | -0.370 | -0.294 | 0.286 | 0.288 |
| 7 | 0.384 | -0.419 | -0.426 | 0.473 | 0.364 | -0.377 | -0.385 |
| 8 | 0.450 | -0.392 | -0.442 | 0.567 | 0.450 | -0.477 | -0.482 |
| 9 | 0.426 | -0.387 | -0.417 | 0.504 | 0.430 | -0.406 | -0.411 |
| 10 | -0.425 | 0.369 | 0.460 | -0.557 | -0.461 | 0.461 | 0.471 |
| 11 | 0.212 | -0.502 | -0.377 | 0.344 | 0.255 | -0.296 | -0.306 |
| 12 | 0.428 | -0.425 | -0.469 | 0.507 | 0.428 | -0.366 | -0.370 |
| 13 | -0.465 | 0.377 | 0.472 | -0.541 | -0.534 | 0.480 | 0.486 |
| 14 | -0.350 | 0.386 | 0.395 | -0.504 | -0.392 | 0.389 | 0.396 |

Table 5.4: Perceptual Test 2D PCA (First Factor) versus Features: Highlights of Lesser Correlations

concluded that 2-3 fundamental axes of simple acoustical form are adequate for describing timbre is that they were dealing with a timbre space of much more limited scope.

Limitations of timbral scope and lack of many data points can be explained in that manner. The following subsections show that the feature set detailed here is adequate for explaining the timbral nuances in the set of 153 sounds. Thus the low number of axes solutions considered in this subsection are in fact masking levels of nuances (a complex dimensional form) in a superficially simple structure.

5.6.2 3D Hierarchical Decomposition of Timbre Space

Using the AXESDIST program, it is possible to investigate the 335-dimensional data space to understand better the sort of spectral features which relate to particular timbral group differences in the set of 153 sounds. The group sets used in this part of the analysis are detailed in Table 5.5. Some of the groups are similar to those used in Chapter 3 for displaying the relationships resulting from the PCA of the perceptual tests (see Table 3.5).

| Group Name | Stimulus Codes |
|-----------------|---------------------------------|
| Strings | 1-27 |
| Woodwind | 28-51 |
| Brass | 52-67, 151 |
| Hammered Tonal | 68-79, 81-83 |
| Percussion | 84-109 |
| Synthetic A | 111-130 |
| Synthetic B | 131-145 |
| Open Strings | 2,4,6,8,10,11,12,15,16,17,20,22 |
| Stopped Strings | 1,3,5,7,9,13,14,18,19,21,23,24 |
| Pianos | 68-75 |
| Hammered Misc | 76-79,81-83 |

Table 5.5: Groups Used in Subsection 5.6.2 for AXESDIST Analyses

The method employed in this subsection is based on hierarchical decomposition of the timbre space, to investigate the data structures at different levels of detail and from different perspectives. As is shown in the first part (5.6.2.1) below, trying to precisely decipher the relationships between the stimuli from a single level in a manner that can be visually represented is not viable. However, much more can be understood by progressively extracting timbral forms from the overall set. 3D result sets are extracted as they are the largest dimensional form that can be easily graphically represented on paper, but other numbers of dimensions could have been used. Indeed, the optimal number of dimensions in any particular circumstance may be greater or fewer than 3, but the time for such extensive analysis was not available to the author. Certain combinations of methods are utilised in the following parts, to give a set of related viewpoints on the timbral distinctions being analysed. Many others could have been used given more time, but those listed are a perfectly valid representative structure.

5.6.2.1 High-Level Distinctions

The arrangement of “traditional” groups in the structure of perception was examined in Chapter 3, and in particular in Subsections 3.5.4 and 3.5.5. A similar structure can be investigated in AXESDIST with the minimisation equation as follows:

$$\begin{aligned}
& \text{Within}(\text{Strings}, n\text{Method}) + \text{Within}(\text{Woodwind}, n\text{Method}) \\
& + \text{Within}(\text{Brass}, n\text{Method}) + \text{Within}(\text{HammeredTonal}, n\text{Method}) \\
& \quad + \text{Within}(\text{Percussion}, n\text{Method}) \\
\hline
& \text{Between}(\text{Strings} \leftrightarrow \text{Woodwind}, d\text{Method}) + \text{Between}(\text{Woodwind} \leftrightarrow \text{Brass}, d\text{Method}) \\
& + \text{Between}(\text{Brass} \leftrightarrow \text{HammeredT}, d\text{Method}) + \text{Between}(\text{HammeredT} \leftrightarrow \text{Percussion}, d\text{Method}) \\
& \quad + \text{Between}(\text{Percussion} \leftrightarrow \text{Strings}, d\text{Method})
\end{aligned} \tag{5.6}$$

The results of AXESDIST are that the best minimisations are achieved with the feature triples given in Table 5.6. This details all combinations of methods in the simplest all equal methods scheme. It is noticeable that these show distinct similarities to the high correlating results of Subsection 5.6.1. That is, at a broad level it can be seen that the results of searching for the axes which distinguish between the traditional groups has similarities to the perceptual structure correlations with the acoustical feature set. Feature 19.0 (mode 0 of amplitude envelope) has a strong presence, as it did in the correlations. It and the other features relate strongly to the results of previous researchers in finding major axes; such as the onset portion and variation in the onset (25.3RO, 25.6RO, 11.7RO), frequency balance (25.3RO, 24.0AA, 25.6RO, 26.0SA, 24.0AR, 21.0RR) and harmonic form (3.0AR, 11.7RO). Again, similarities can be seen with the Grey results ([71]) among others.

| <i>Methods Used</i> | <i>Resulting Features</i> |
|---|---------------------------|
| <i>nMethod=averageall, dMethod=averageall</i> | 19.0SF, 25.3RO, 25.6RO |
| <i>nMethod=averageall, dMethod=mindistall</i> | 3.0AR, 19.0SF, 24.0AA |
| <i>nMethod=averageall, dMethod=centroiddist</i> | 19.0SF, 25.3RO, 25.6RO |
| <i>nMethod=maxdistall, dMethod=averageall</i> | 19.11SF, 20.2SF, 26.0SA |
| <i>nMethod=maxdistall, dMethod=mindistall</i> | 3.0AR, 19.0SF, 24.0AR |
| <i>nMethod=maxdistall, dMethod=centroiddist</i> | 11.7RO, 19.0SF, 21.0RR |

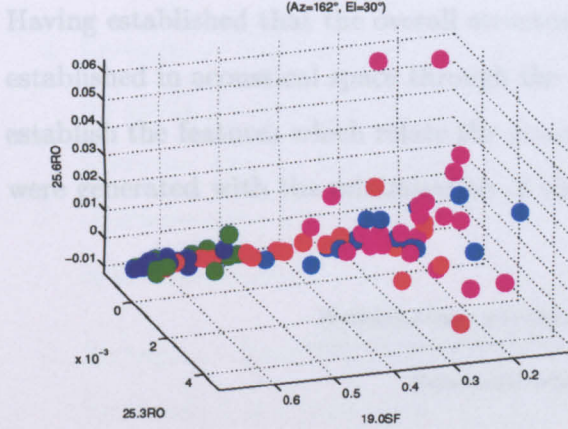
Table 5.6: Results for AXESDIST Minimisation of Equation 5.6

The large dot plots corresponding to the results of Table 5.6 are given in Figure 5.1. These plots display distinct similarities in structure to those derived in Subsection 3.5.5 from the perceptual results (but note that the colouration is slightly different). These similarities are interesting as the results are derived from separate viewpoints on the stimuli. Note that there is clustering and mingling of the groups which the AXESDIST program has tried to separate, which has similarities with the results in Chapter 3. The author believes

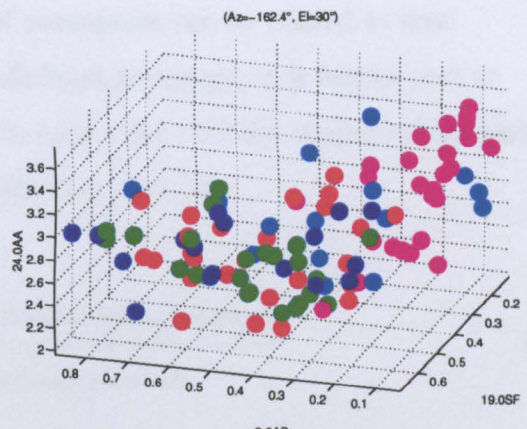
that this hints that the acoustical information in the stimulus set interpreted as a 3D space (and partially limited by the size of the feature space) is inadequate to describe all the differences involved. Those differences can be perceived - woodwinds do not sound as close to brass as they appear in the plots for example. Therefore, it is the assumption that a 3D space is adequate for describing all the differences in the stimulus set that is at fault.

The different *nMethod* and *dMethod* combinations produce different biases in the minimisation scheme used in AXESDIST. It is incorrect to draw conclusions concerning the effectiveness of particular methods in general from a single minimisation scheme. In a different situation the least effective scheme here could be the most effective. The most important point is that none of these is the “right” answer; they represent particular objective views of the most appropriate features in 3D space based on the particular minimisation and distance scheme.

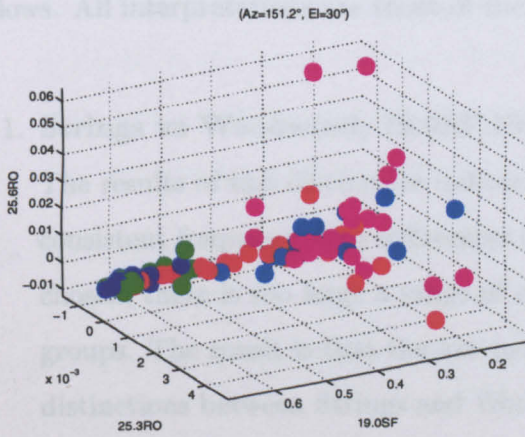
5.6.2.2 Instrument Group Distances



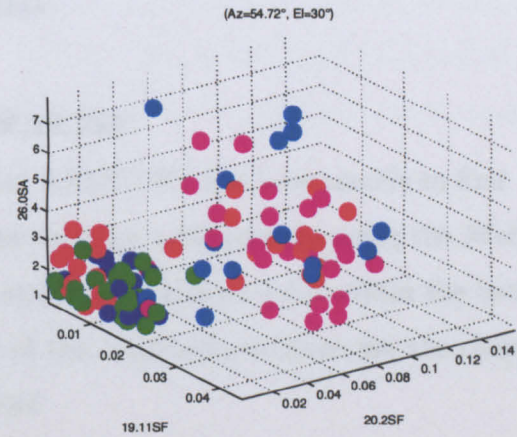
(1) $nMethod=averageall, dMethod=averageall$



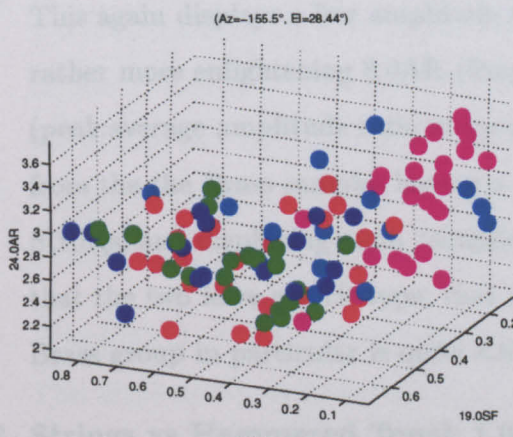
(2) $nMethod=averageall, dMethod=mindistall$



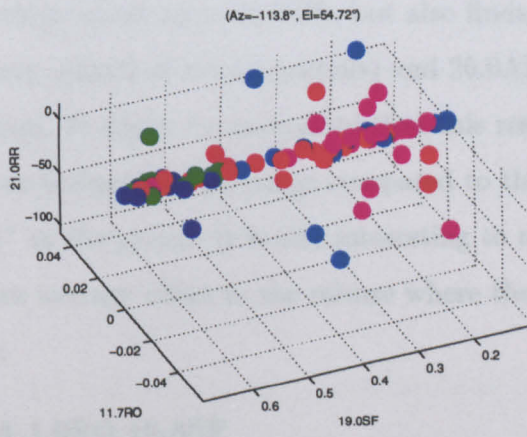
(3) $nMethod=averageall, dMethod=centroiddist$



(4) $nMethod=maxdistall, dMethod=averageall$



(5) $nMethod=maxdistall, dMethod=mindistall$



(6) $nMethod=maxdistall, dMethod=centroiddist$

Figure 5.1: Large Dot Plots for AXESDIST Results for Equation 5.6 (red=Strings, green=Woodwind, blue=Brass, cyan=Hammered Tonal, magenta=Percussion)

5.6.2.2 Instrument Group Distinctions

Having established that the overall structure of perception can be related to that established in acoustical space through the traditional groupings, it is logical next to establish the features which relate the groups to each other. All the results in this part were generated with the minimisation of the following equation:

$$\frac{\textit{Within}(\textit{Group}_1, n\textit{Method}) + \textit{Within}(\textit{Group}_2, n\textit{Method})}{\textit{Between}(\textit{Group}_1 \leftrightarrow \textit{Group}_2, d\textit{Method})} \quad (5.7)$$

The *nMethod* used is *maxdistall* and the *dMethod* is *centroiddist*. The results are as follows. All interpretations are those of the author:

1. Strings vs Woodwind; 19.0SF,19.1SF,19.2SF

The results of this distinction indicate that AXESDIST has been unable to find consistent frequency-type differences in the form for a 3D solution using the *Methods* chosen; there is too large a range of non-amplitude envelope form within the two groups. The result is that the low modes of the amplitude envelope are the clearest distinctions between Strings and Woodwind.

2. Strings vs Brass; 8.0AR,19.0SF,26.0AR

This again displays a low amplitude envelope mode axis (19.0SF), but also finds a rather more enlightening 8.0AR (frequency spread of strong partials) and 26.0AR (peak:average amplitude ratio in spectrum). It might be speculated that this results from the the Brass samples having a more limited timbral range compared to the Strings, and displaying more “simplicity” in the group. It is also interesting to note that the two have an AR type; that is, an average effect in the release where the Brass group in particular is quite stable.

3. Strings vs Hammered Tonal; 1.0AA,1.0SO,19.8SF

Here the distinction produces effects which concentrate on the more dynamic elements of the sounds, compared to the Strings vs Brass results. Feature 1 is the proportion of harmonic partials, which can be expected to be lower in the Hammered Tonal set, both throughout the sound (1.0AA) and in its varying nature in the onset (1.0SO). Feature 19.8 also shows the dynamic nature distinction between the HT and

Strings. The picture is necessarily confused by the timbral similarities between HT and some of the String sounds as discussed in Subsection 3.5.3.

4. Strings vs Percussion; 1.0SA,1.0AR,26.0AO

As might be expected, this result set has similarities to the Strings vs HT one. However, 26.0AO has replaced 19.8SF, as the Percussion group has more “noise” characteristics than the Strings set, leading to a difference in the peak:average level in the frequency profile.

5. Strings vs Synthetic A; 16.0SF,17.0SF,25.1AR

Synthetic A is a particularly strange group as it is not an homogeneous timbral form in the sense of being an easily imagined family of sounds. But it is a timbral group in displaying “unnatural” characteristics. They were mostly found to be distinctively synthetic in the perceptual test (Subsection 3.5.3). The AXESDIST results reflect that; features 16 (number of peaks and troughs) and 17 (breakpoints) both concern the “complexity” of the amplitude envelope.

6. Strings vs Synthetic B; 19.0SF,19.1SF,19.2SF

This result sees the “complexity” issue from a different perspective; that is, the extreme modal aspects of a plain rectangular envelope which characterise the test tones.

7. Woodwind vs Brass; 1.0RA,11.3AO,11.5SO

Woodwind and Brass have known similarities in timbral form, so the distinguishing characteristics are more subtle than the broad result axes of these two groups against the Strings set. The classical wood-like characteristic of 3rd and 5th harmonics is present (11.3AO, 11.5SO). The other feature concerns the variation of harmonic proportion over the length of the sound.

8. Woodwind vs Hammered Tonal; 1.0RA,19.0SF,21.0RR

This displays the same feature of variation of harmonic proportion (1.0RA) as with the Woodwind vs Brass distinction. However, the other two features relate to the more percussive nature of the HT set compared to the Woodwinds, with the amplitude mode 0 offset and slope of the frequency profile.

9. Woodwind vs Percussion; 10.0AO,10.0AR,19.0SF

As with *Woodwind/HT*, 19.0SF is present, but it is interesting to have two features relating to the lower frequency of the strong partials in the onset and release. It is logical, as the percussive stimuli in general have less bass presence, but interesting that it is stronger than other features such as other measures of amplitude envelope shape.

10. **Woodwind vs Synthetic A; 2.0SA,8.0AR,20.5SF**

This is very much harder to interpret than the *Strings versus Synthetic A* distinction. It could be that the few string-like sounds in the *Synthetic A* group and its significant range of timbral form forced the results against the *Strings* set into the “complexity” aspects that might be expected to distinguish *Synthetic A* from the others. Here though, the proportion of strong inharmonics (2.0SA), the frequency spread of strong partials (8.0AR) and a low frequency mode (7.5Hz) of the amplitude envelope (20.5SF) are important.

11. **Woodwind vs Synthetic B; 11.2SO,11.8AR,19.0SF**

Rather than being characterised mainly by the plain shape of the *Synthetic B* group, as with *Strings vs Synthetic B* and others later, this distinction has the emphasis more on the character of the tones; the variation of harmonic 2 in the onset (11.2SO) and the average level of harmonic 8 in the release (11.8AR). However, the mode 0 level still makes an appearance.

12. **Brass vs Hammered Tonal; 1.0RA,19.0SF,21.0RR**

This is the same result as *Woodwind versus Hammered Tonal*, for the reason that *Woodwind* and *Brass* are perceptually tightly coupled.

13. **Brass vs Percussion; 19.0SF,20.0SF,20.2SF**

This is simpler than *Woodwind vs Percussion* and is the sort of result that would have been predicted in advance for that and this distinction; low mode/frequency axes.

14. **Brass vs Synthetic A; 9.0RA,21.0AR,23.0RA**

Again difficult to interpret, this result does show a similarity with *Woodwind/Synthetic A* in that the upper frequency of strong partials (9.0RA) has similar background to 8.0AR. Also, the variation in centroid position (23.0RA), frequency slope (21.0AR) and variation in the proportion of strong inharmonic

partials (2.0SA, in Woodwind distinction) are influenced by the amount of noise-like energy which is more prevalent in the non-synthetic sounds.

15. Brass vs Synthetic B; 9.0SO,19.2SF,19.3SF

This distinction has two amplitude envelope shape characteristics, similar to Strings vs Synthetic B, because of the plain nature of the Synthetic B envelope. The variation of upper frequency of strong partials axis may relate to the breathy nature of brass onsets (9.0SO).

16. Hammered Tonal vs Percussion; 2.0AA,2.0AO,2.0AR

This result is very strong indication of the influence of the proportion of strong inharmonic partials. Given the similarities of the two groups, it is the presence of the inharmonic partials in the overall form (AA,AO,AR) which is most important.

17. Hammered Tonal vs Synthetic A; 9.0RA,23.0RA,25.3RR

The similarities with Brass versus Synthetic A with 9.0RA and 23.0RA are apparent. The influence of the 3rd split bin (25.3RR) is not as easily explained by the “noisiness” theory.

18. Hammered Tonal vs Synthetic B; 11.10RO,19.2SF,19.3SF

Similarities in envelope form features (19.2SF, 19.3SF) with the other distinctions concerning Synthetic B are again apparent. But, again the other result axis relates to the tone of the sounds being different; here it is the 10th harmonic in the onset.

19. Percussion vs Synthetic A; 6.0AA,6.0SR,20.17SF

Again, this is hard to interpret. The noise properties are apparent through the inharmonic/harmonic proximity (6.0AA,6.0SR) and the low frequency variation in the envelope (20.17SF, 25.5Hz), however.

20. Percussion vs Synthetic B; 14.0SF,19.0SF,19.2SF

These again relate strongly to amplitude envelope. This is shown with other distinctions concerning the Synthetic B set, but is particularly easy to interpret with the Percussion group where the amplitude attack to average ratio, and two low amplitude envelope modes are the results.

21. Synthetic A vs Synthetic B; 9.0SO,19.2SF,19.3SF

The subtleties and problems with Synthetic A in other distinctions are less apparent here, where the strong difference is again the amplitude envelope form. The upper

frequency of strong partials in the onset also makes an appearance, as it did with the Brass/Synthetic B distinction

The author's interpretation of the results above is based mainly on treating them as isolated axes, when (as can be seen from the plots of Figure 5.1) it is the set of result axes *together* which produces the particular spread which is found to minimise the AXESDIST equation. As such, interpretation in this manner is not always totally clear due to the interactions between acoustical components in the solution.

5.6.2.3 Lower Level Group Distinctions

Below the level of the general instrument group distinctions considered above, it is possible to consider the differences between any number of timbral subspace groups. What limits informative investigation is the number of instances of related timbral types that can be examined. The set of 153 stimuli used in this research does not present a comprehensive enough data set to effectively investigate, say, the nuances of woodwind technique, as there are not enough examples of each instrument. However, it is still possible to consider timbral distinctions below the instrument group level, such as the examples below. These use Equation 5.7 again and with $nMethod = \text{maxdistall}$ and $dMethod = \text{centroiddist}$. The interpretations are those of the author:

1. Open Strings vs Stopped Strings; 6.0AA,6.0AR,11.10SR

Both the open and the stopped strings groups have similar pitch range and timbral forms represented within them. As such, the distinguishing features are looking for nuances of difference relating to the stopped/open distinction. Feature 6 is inharmonic to harmonic partial proximity, possibly relating to the ringing nature of the open strings. 11.10SR is a measure of the variation of the 10th harmonic in the release portion. This is an hard result to explain; but it is also one that would be very hard to predict in advance or to spot by hand-examination of the feature data.

2. Pianos vs Hammered Misc; 7.0AA,10.0AA,10.0AR

Here the general shape of the amplitude curves being examined is similar, but the frequency-domain differences are more apparent. Thus the results include the frequency spread of harmonics (feature 7) and the the lower frequency of strong

partials (feature 10).

These are just a couple of examples of a large number of potential distinctions which could be examined. As the distinctions focus more on nuances, it becomes harder to predict and explain the results. This indicates the importance of an objective approach to feature extraction and analysis. Also, as the analysis focuses more on nuances the effect of the chosen timbre space becomes more important as the particular features and stimuli chosen are being more closely examined and there are fewer data points being considered. This makes the results less widely applicable.

5.6.2.4 General Notes on Hierarchical Decomposition

It is difficult to draw accurate broad conclusions based on the results described above concerning the acoustic components which characterise particular instrument groups. What is apparent to the author, though, is the large range of features in the results of the hierarchical decomposition. This shows that it is not reasonable to consider timbre a low dimensional aspect of sound. It has “strong” elements, which are consistently found to be important in distinctions, which could be classed as the major axes which describe the timbral structure. But a much larger number of features are relevant which also fill in the less prominent perceptual details.

As regards the fundamentally important axes of all timbre, there is no “right answer” at present. Different methods achieve slightly different resulting feature sets. Eventually researchers may find that all the different perspectives can be best represented in a single unified form which takes in all the views gained from different methods of searching the data space. For now, though, each objectively generated view is a valid one as it indicates a certain optimised way of looking at timbral distinctions.

The experiments show how an hierarchical decomposition of timbre space enables a more structured analysis of the data space than is possible by considering a single overall level through a plain multidimensional form. With a large data set, being able to focus the scope of the consideration aids in interpretation. Future research may be able to develop more composite features which can be used to explain parts of these distinctions, and others, more simply, with the aim to forming SCAs (Section 2.4).

5.6.3 Investigation of Dimensional/Noise Effects

It is not obvious what number of timbral dimensions is likely to best represent timbre as a whole, or any particular timbral distinction. Grossly different timbral groups are likely to be distinguished by a single simple acoustical feature. But subtly different groups may require many more such features. The situation can be made more complex by considering more carefully weighted schemes than simply distinguishing two groups; such as the perceptual ordering of similarities within the groups being distinguished, rather than just the axes which differentiate between the groups as a whole.

As discussed in Subsection 5.5.2, the intention is to find the number of features required to perceptually order some degrees of timbral difference. Eventually timbre may be described in terms of perceptually strong, logically-organised fundamental dimensions. Any timbral distinction could, then, be described in those axes. This research necessarily starts from a less composite set of features, which are generally simple acoustical forms known to relate to timbral perception. The eventual aim of timbre research is, rather, *perceptual* features represented in (potentially complex) acoustical forms. Because the features used here are generally simple, it might be expected that the dimensionality required to best represent timbral distinctions here will be larger than that achievable with more composite acoustical forms. However, that doesn't mean that the fundamental dimensions of timbre space are as low as 3-5 in number as estimated by previous authors ([77], for example).

It is of interest to find the sort of dimensionality present in the current feature/timbre space. What is apparent from the author's analysis of searching the feature set for particular timbral distinctions is that there are a number of components at work:

1. The size and scope of the feature set. If the feature set is small then necessarily the fit of individual features in providing the best solution in distinguishing between different timbral forms *among the available group* appears to be good. For example, if assessing spectral centroid, there will be a number of situations where it appears to provide reasonable fit in distinguishing groups of sounds. However, if a number of related features are also considered, then it may be found that different features are more appropriate than others in different situations. Thus, there are hidden dimensions which improve the fit of the features to the distinctions. Overall, the effect is that only by considering a wide range of features at once can a reasonable

assessment of the actual dimensionality of the situation be achieved.

2. The size and scope of the stimulus set. A small number of sounds can be effectively distinguished by a small number of acoustical features. Thus it can appear that 3-5 dimensional solutions can distinguish all the timbral differences between a small set of sounds effectively. If however there are a large number of stimuli with complex similarities and scope then a small feature set of simple acoustical measures can no longer produce a perceptually-structured model.
3. Features which contribute to improved distinction. In a particular situation where a timbral distinction is being investigated, two types of positive contribution from features are likely: Firstly, individual features may be good at perceptually ordering the stimuli being considered in the manner of interest. But also, groups of features may work together in creating distinctions to which the features individually contribute less.
4. Features which contribute nothing to distinction. When searching the data space for features which improve a distinguishing metric, an algorithm such as AXESDIST will find features which do not *harm* the distinction but do not actually do any good either. This occurs naturally as part of the searching process. These artificially raise the perceived dimensionality of the distinction unless the results are carefully considered.
5. Features which contribute noise. These are features which when chosen only harm the distinction. If the distinguishing metric starts to rise with higher numbers of features, then noise is being added; the dimensionality is not actually as high as this.

These effects can be demonstrated by considering different distinctions using the AXESDIST program. The indicator of success in the minimisation is the value of the ratio which results from the algorithm; that is, the smallest quotient of the distances within the groups to those between the groups for a particular number of output dimensions. This value can be compared between different numbers of solution dimensions to find the most appropriate degrees of freedom to explain the timbral difference.

5.6.3.1 High Level Distinction Example

In this example, the same minimisation form is used as in 5.6.2.1:

$$\begin{aligned}
 & \text{Within}(\text{Strings}, \text{maxdistall}) + \text{Within}(\text{Woodwind}, \text{maxdistall}) \\
 & + \text{Within}(\text{Brass}, \text{maxdistall}) + \text{Within}(\text{HammeredTonal}, \text{maxdistall}) \\
 & + \text{Within}(\text{Percussion}, \text{maxdistall}) \\
 \hline
 & \text{Between}(\text{Strings} \leftrightarrow \text{Woodwind}, \text{centroiddist}) + \text{Between}(\text{Woodwind} \leftrightarrow \text{Brass}, \text{centroiddist}) \\
 & + \text{Between}(\text{Brass} \leftrightarrow \text{HammeredT}, \text{centroiddist}) + \text{Between}(\text{HammeredT} \leftrightarrow \text{Percussion}, \text{centroiddist}) \\
 & + \text{Between}(\text{Percussion} \leftrightarrow \text{Strings}, \text{centroiddist})
 \end{aligned} \tag{5.8}$$

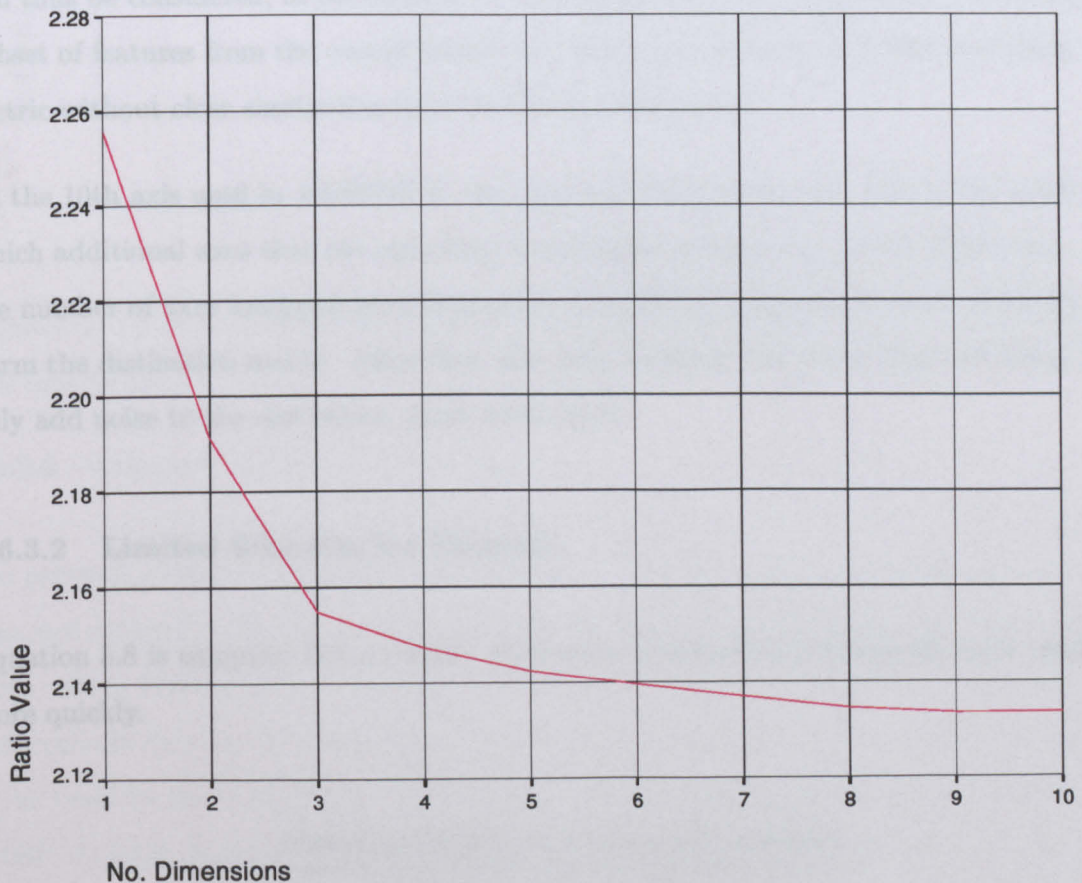


Figure 5.2: Values of Minimisation Ratio for Equation 5.8 Against Number of Dimensions; Full Feature Set

This is a complex timbral distinction, which is predictably hard to characterise well with only a couple of features from the set in use. Figure 5.2 shows the minimised result ratios

for the equation which AXESDIST produces. Up to 9 axes the ratio improves (i.e. decreases), and then the ratio levels off which shows that adding more dimensions is failing to improve the ratio. To understand this result it is necessary to examine the features which compose the answer. The lowest ratio is achieved with 9 features, which are 1.0RO (proportion of all partials which are harmonic), 10.0AA (lower frequency of strong partials), 10.0AR, 11.5RO (harmonic 5), 11.7RO (harmonic 7), 19.0SF (mode 0 of amplitude envelope), 21.0RR (slope of spectrum), 25.4RO (amplitude in split frequency bin 4 of 8) and 25.6RO (amplitude in split frequency bin 6 of 8). These cover a large range of aspects of the time-varying frequency spectrum form. If they had all been very similar it would have been unreasonable to suggest that 9 axes is a correct assessment of the dimensionality necessary to effectively distinguish between the groups. 10.0AA and 10.0AR are very similar, so 8 may be a more logical conclusion, though. Dimensionality can thus be considered, in the context of an algorithm such as AXESDIST, as the largest subset of features from the overall feature set which produces the best distinguishing metric without close similarities between the features chosen.

At the 10th axis used in AXESDIST, the graph of ratios levels out. This is the point at which additional axes that the algorithm found add nothing extra to the distinction. As the number of axes increases from that point, AXESDIST chooses the axes which do not harm the distinction metric. After that, the ratio starts to rise as the features being added only add noise to the distinction (least noise first).

5.6.3.2 Limited Stimulus Set Example

Equation 5.8 is complex, but a simpler distinction reaches the levelling-off point much more quickly.

$$\frac{\textit{Within}(\{1,72\},\textit{maxdistall})+\textit{Within}(\{2,73\},\textit{maxdistall})}{\textit{Between}(\{1,72\}\leftrightarrow\{2,73\},\textit{centroiddist})} \tag{5.9}$$

The value of the AXESDIST minimisation ratio for Equation 5.9 is plotted in Figure 5.3. Increasing the number of dimensions from 1 has no positive effect on the minimisation. The small distinction of the equation is adequately covered by a single dimension.

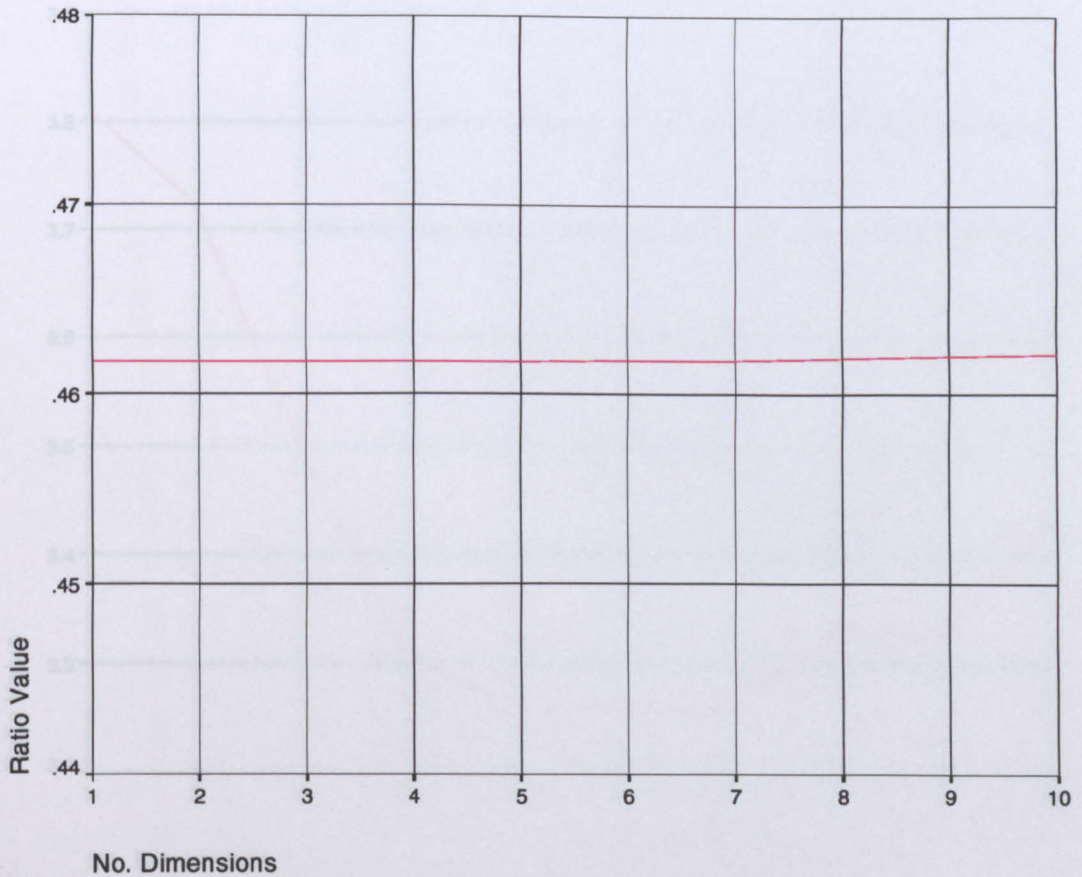


Figure 5.3: Values of Minimisation Ratio for Equation 5.9 Against Number of Dimensions; Full Feature Set

5.6.3.3 Limited Feature Set Examples

The previous example shows that with small numbers of stimuli in a distinction, the apparent dimensionality is small. A situation which also limits apparent dimensionality exists with a limited feature set. The examples that follow use Equation 5.8 again. The first example uses the following feature subset (rather than the full set as used previously) as the basis for the AXESDIST minimisation: 3.0AO, 3.0AR, 8.0AO, 8.0AR, 12.0SF, 14.0SF, 21.0AO, 21.0AR, 23.0AO, 23.0AR. This set is a broad range of features, also taking in the differences between the onset and release portions.

Figure 5.4 shows the minimised result ratios which AXESDIST produces. Compared to the results of Figure 5.2 the apparent dimensionality has been reduced from about 8 to about 5 or 6. This reduction is curtailed by the fact that the equation being minimised has

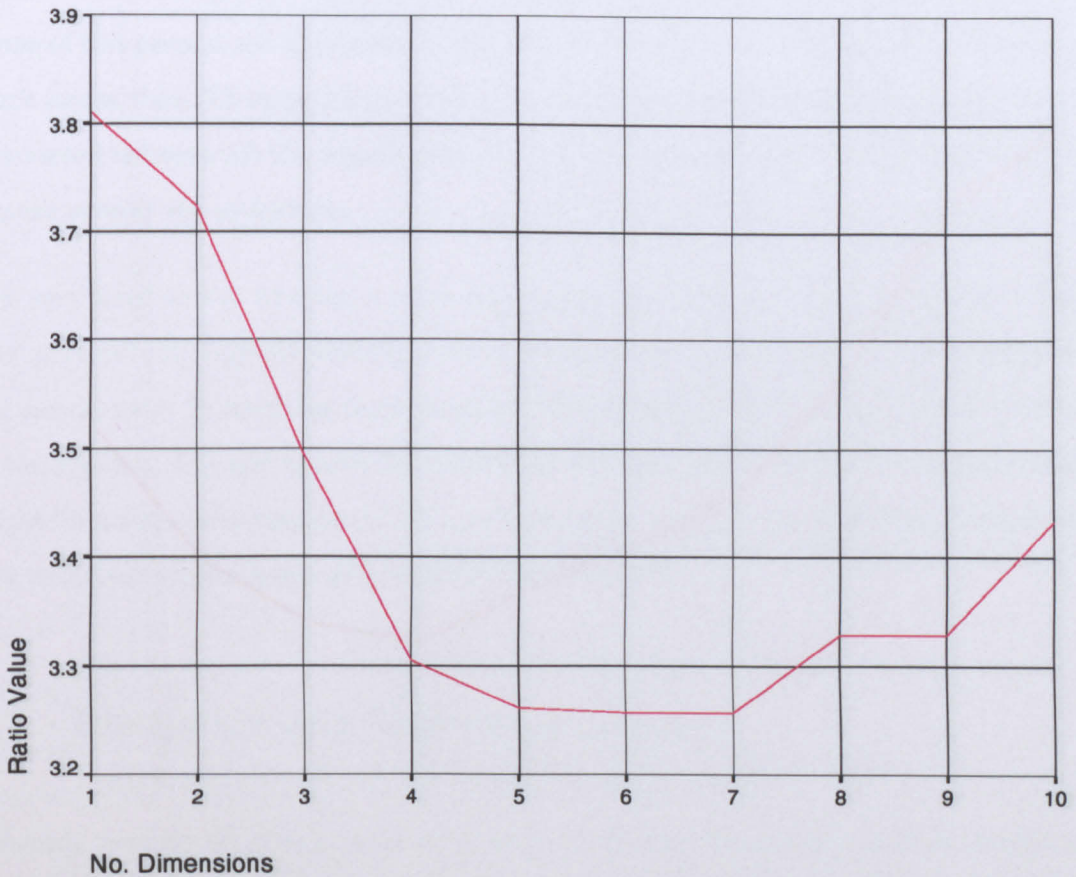


Figure 5.4: Values of Minimisation Ratio for Equation 5.8 Against Number of Dimensions; First Limited Feature Set

a broad range of characteristics and the features used to represent it have been deliberately chosen with a broad range. Thus the features actually still work quite well.

The second example uses the rather more limited feature subset of 1.0AA, 7.0AA, 9.0AA, 11.2AA, 11.3AA, 11.4AA, 11.5AA, 11.6AA, 11.7AA, 11.8AA. This subset concentrates on aspects relating to the harmonics of the stimuli as an average over the whole sound. Figure 5.5 shows the minimised result ratios which AXESDIST produces for this set. This time, the more limited range of features trying to describe a broad timbral form results in a best fit at 4 axes. In part this shows the importance of harmonic form in describing timbral distinctions. However, it is half the number of features found to be most appropriate for describing the same distinction based on the whole feature set.

Limited stimuli in the distinction (in 5.6.3.2) and a limited feature set both result in low dimensional results. From the point of view of a timbre space being limited by the features

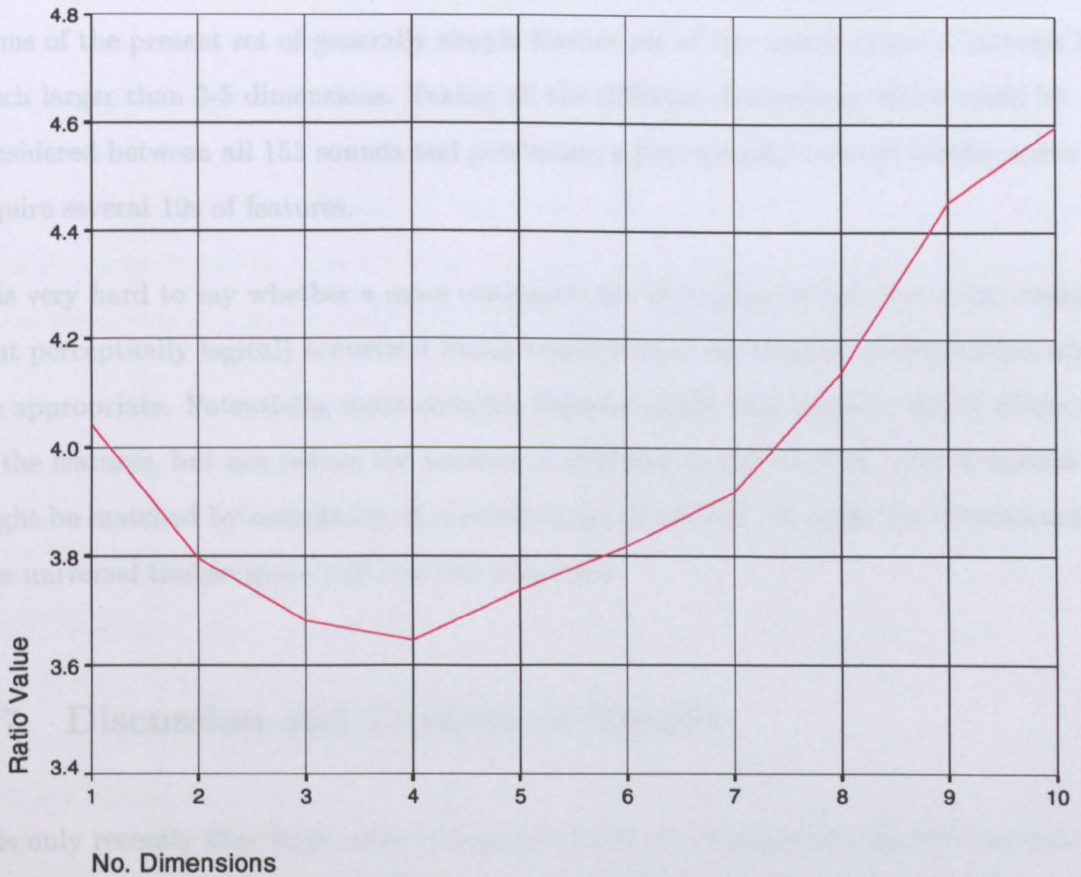


Figure 5.5: Values of Minimisation Ratio for Equation 5.8 Against Number of Dimensions; Second Limited Feature Set

and stimuli that describe it, that is not a problem. However, when looking to extrapolate the results to a wider context, a limited feature set gives a false impression of the actual dimensionality which exists. Also, the limited features leads to false conclusions as to the nature of a distinction as the statistical methods can, necessarily, only choose from the features that are available.

5.6.3.4 Dimensionality Conclusions

Previous research has often been restricted in both stimulus set and feature space which necessarily leads to a limited timbre space. Quite reasonably, those authors have concluded that their experiment shows a dimensionality of 3-5. However, a problem arises when trying to predict the dimensions of a universal timbre space. Previous research has often, seemingly, underestimated the complexity of timbre perception. The experiments of

Subsections 5.6.1 and 5.6.2 as well as the present one show that the degrees of freedom in terms of the present set of generally simple feature set of the timbre space of interest is much larger than 3-5 dimensions. Taking all the different distinctions which could be considered between all 153 sounds and producing a perceptually ordered feature space may require several 10s of features.

It is very hard to say whether a more composite set of features which used more complex (but perceptually logical) acoustical forms would reduce the number of dimensions which are appropriate. Potentially, more complex features might only improve the fit of the data to the features, but not reduce the number of different facets involved. Any reduction might be matched by complexity of a wider range of stimuli. As such, the dimensionality of a universal timbre space still remains unknown.

5.7 Discussion and Context of Results

It is only recently that large scale evaluation of the use of extracted spectral features in distinguishing a wide range of stimuli has become viable with personal computers. Comparing the results here to similar studies is hard as this study considers a considerably different timbre space to other research. The collected interpretations of previous researchers as detailed in Sections 2.10 and 2.11 provide information to guide new research. But with bigger timbre spaces being considered, new information emerges as to the complex hidden acoustical features which affect timbre perception. Even so, it is apparent from the results in this chapter that those parameters as derived by previous researchers can often be supported to a degree in different circumstances. This doesn't prove that those researchers were exactly right, but it does indicate the potential for finding universal descriptions in future. It is hoped that even if the timbre spaces used by future researchers are somewhat different, similar conclusions may be drawn concerning the areas of investigation considered here.

Information and knowledge are not the same thing. A vast number of measures could be extracted from a spectral form. Not only might such measures be quite correlated, neither do they represent a proportional degree of knowledge through their uncorrelated aspects. The increase of knowledge fundamentally requires the use of techniques to analyse the metrical relationships between the features to achieve a set of solutions. Such solutions

require comparison and interpretation, which means that there is no such thing as “the solution”. However, each solution indicates the sort of features which fit with the perceptual area of interest, through which a picture of timbral perception can be built up. The chosen feature set in this chapter is effective because it has a considerable range of acoustical forms which relate well to known relevant aspects of the time-varying spectrum. A larger range would be interesting to consider, but the results show logical structure and prove the points of interest within the timbre space considered.

Langmead suggests that mechanisms in the brain attempt to match parametrically ([107]). It is interesting to speculate whether the auditory cortex extracts features in manners similar to those described here. It is likely that the hearing system has a natural ability to evaluate the timbral form of stimuli, but can also develop new mechanisms ([89]). The major “standard” feature detections might be those that are consistently found to be important by researchers. However, the author suggests that research has not yet become sophisticated enough to conclusively explain even the major features, let alone more subtle analysis which might be developed by individuals as they learn about the sonic World.

This study considers a much larger timbre space and in more depth than previous research. This chapter provides a set of methods which can be expanded upon and compared in future work. The sort of expansions which might be considered include techniques which can cope with different contexts of acoustic form more readily, such as different environmental effects. Also, there may be a limit as to how far it is possible to reasonably expand the number of features and stimulus inputs and gain adequate discriminatory information from a single sample. Potentially, a number of stimuli may be required to trace out an area of the instrument’s subspace for correct identification.

5.8 Conclusions

Compared to the studies listed in Section 2.8 and other relevant material, the work outlined in this chapter is novel in the following ways:

1. The experiments use a considerably greater number, range and detail in the stimulus set than has been previously used, and can thus claim to be more comprehensive and universally applicable than previous work.

2. The feature extraction and analysis study is based on a much larger acoustical feature set than has been previously examined.
3. The feature searching algorithm (AXESDIST) uses an especially designed exhaustive enumeration technique to examine the feature set in a more appropriate manner than is possible with the standard scaling and discrimination techniques used by previous researchers.
4. This chapter includes an empirical examination of dimensionality and hierarchical decomposition of acoustical timbre space with relation to perception, which has not been attempted before.
5. The study is the most comprehensive single examination of the relationship between the spectral/acoustical form and timbre perception to date through its use of a number of related experiments.

With reference to the original aims of this chapter outlined at the end of Section 5.2, this feature extraction and analysis study demonstrates the following points:

1. Features can be extracted from the spectral form which enable sound qualities to be distinguished; and those features are not arbitrarily linked to perceptual differences, but show logical relationships. The feature extraction algorithms of Section 5.4 achieve objective measurements of spectral parameters by avoiding user intervention and provide data which was found to logically relate timbral forms in Section 5.6 through the objective statistical mechanisms employed. This shows that the acoustical features (developed from previous research in Sections 2.9 and 2.10) can be linked to the structure of timbre perception. Moreover, logical relationships are apparent, such as the correlations between amplitude envelope modes and the first PCA axes in Subsection 5.6.1, and the features found to relate to timbral distinctions in 5.6.2.2, and so on.
2. Hierarchical discrimination of sound qualities is an effective system for structured decomposition of timbre space. Through the experiments of Subsection 5.6.2 it is apparent that if timbre space is treated as a group of embedded distinctions, it is possible to consider the data set in manageable units. While it is difficult to comprehend the relationships between the stimuli at the highest level in a precise

manner, focusing on particular distinctions in an hierarchy allows the researcher to understand the acoustical mechanisms at work in structured ways. Many previous researchers have failed to appreciate this, it seems.

3. The dimensionality of timbre space is considerable and the perceived differences between all perceptible stimuli cannot be completely explained by a low number of dimensions. Previous research has often attempted to explain the timbral differences in a small stimulus set by a plain multidimensional form and succeeded in achieving moderate fit with 3-5 dimensions (Section 2.7.1). To explain the timbral structure with many more stimuli (such as the 153 use in this research) and with greater regard to accuracy in explaining the nuances of perceptual difference requires many more dimensions. Dimensionality is the best set of axes chosen from the feature set for distinguishing between groups of stimuli in a perceptually structured manner, as explained in Subsection 5.5.2. The experiments detailed in this chapter show that the number of dimensions of timbre space is large. The correlation experiment in 2D and 3D demonstrates hidden structure which prevents high correlation with the non-principal dimension(s) (Section 5.6.1). The hierarchical decomposition experiment of Subsection 5.6.2 shows how a number of layers of detail are required to adequately explain basic group distinctions, and which has a considerable range of relevant features (i.e. dimensions) in the solution sets. Subsection 5.6.3 discusses the effects of different factors on the dimensional form. Even if a more composite feature set were to be used, the number of degrees of freedom that exists in timbre space as a whole would be found to be considerable.

CHAPTER 6

Conclusions

6.1 Introduction

The research described in this thesis concerns the exploration of the area of auditory perception known as timbre. Both perceptual and acoustical techniques are employed, and the relationship between the two forms is of interest. A set of 153 input stimuli are used in the investigation (see Appendix A). The discussion of the acoustical aspects concerns time and frequency domain representations. Chapter 2 is concerned with previous research in timbre, the different structural forms associated with it, and different definitions concerning timbre and timbre space representations. Chapter 3 is concerned with a study of perceptual similarity judgements between the input stimuli, a statistical analysis of the result structure, and the implications for understanding the structure of timbral audition. Chapter 4 is concerned with analysis and synthesis using a time-varying frequency

spectrum model, with adaptive viewpoint properties to achieve appropriate time-frequency resolution. Chapter 5 is concerned with the extraction of 335 acoustical features from the spectral forms produced by the method of Chapter 4, a statistical consideration to find those features which describe perceptual differences in timbral form between stimuli, and investigation of the associated dimensionality.

6.2 Summary of Chapters

6.2.1 Chapter 2 : Perspectives and Research into Timbre

1. Timbre is an aspect of sound which has been described in a number of different ways, whose complexity and scope limits researchers' ability to neatly classify its nature to the satisfaction of all.
2. The characteristics of timbre with which there is some agreement between authors are as follows (Section 2.3):
 - (a) Timbre studies are concerned with distinguishing between sonic character qualities.
 - (b) Timbre is generally considered to be those aspects of sound which are not pitch, loudness, duration and, possibly, presentation/environment.
 - (c) Timbre is not independent of the other aspects of sound.
 - (d) It is multidimensional and those axes which have been identified to date are neither comparable nor orthogonal.
 - (e) The timbre of sounds can be considered at a number of levels of detail.
 - (f) A particular definition of timbre in research depends upon the scope of the timbre space, which is limited by excluded attributes of sound, the sounds used in the research (their distribution and scope), the model used to describe/analyse/control the timbre space, and the contextual nature of the stimuli.
 - (g) The elements of timbre are often coupled and non-linear, and display continuous qualities.
3. Timbre space is defined by a particular set of research conditions (those used in this research are given in Section 2.4).

4. The number (153) and range (see Appendix A) of sound qualities used in this research are large compared to many previous studies, yet are limited compared to the vast range of audible timbral qualities (Section 2.5).
5. The viewpoint of an investigation into timbre affects the conclusions which can be drawn. Such aspects as whether real-time response is necessary, whether the system is concerned with both analysis and manipulation, and the model structure are important (Section 2.6).
6. The dimensionality associated with the perception of timbre has often been found to be of a low order in previous research, potentially due to the limitations of the timbre spaces concerned (Subsection 2.7.1).
7. There are a number of ways of representing the structure of timbre space. A plain multidimensional set of axes, a set of axes with additional unique components (specificities), and a hierarchical form are typical methods (Subsection 2.7.2).
8. Research techniques generally consider the acoustical and/or perceptual viewpoints. Classical research methods consider evaluation of timbre by semantic scales, by perceptual distance, or by acoustical parameters. All techniques have problems associated with them (Section 2.8).
9. From previous research, the following aspects of spectral form seem to have importance in distinguishing timbral differences; onset/prefix characteristics, steady state, decay, amplitude envelope, harmonic form, inharmonic form, patterning of frequency information, spectral contour, noise aspects, temporal evolution, transient aspects, and synchrony of partial movements (Section 2.9).
10. A number of spectral features have been developed in previous research associated with the spectral aspects listed in the previous point (Section 2.10), but they represent only some of the potential number of features which could be extracted. Furthermore, no single study has considered more than a few features.
11. Spectral features have been correlated with a number of timbral descriptions (semantics) in different contexts (Section 2.11).

6.2.2 Chapter 3 : Perceptual Study

1. There are indications from previous research of a number of mechanisms being involved in timbre perception (Section 3.3):
 - (a) The auditory cortex may process timbre in a number of stages from the low-level form to a number of high-level representations suitable for decision making.
 - (b) Attention may be directed to particular parts of sounds through subconscious (directed by external events) or conscious (directed perception by the listener) focusing.
 - (c) Both categorical and continuous assessment appears to occur in timbre perception.
 - (d) Similarity perception is believed to be organised along dimensional or symbolic lines. It is possible that both may be correct in different parts of the timbre judgement process.
 - (e) Source identification may be based on a framework of relationships to previously experienced sounds.
 - (f) There is little concrete information concerning the effects of prior knowledge on timbral perception at present.

2. The perceptual study in this research uses a technique with the following characteristics (Section 3.4):
 - (a) It is based on template similarity judgements.
 - (b) It is composed of tests where all 153 sounds are rated for similarity to templates of 4 sounds in 6 timbre families.
 - (c) Each sound is rated for timbral similarity on a four point scale.
 - (d) 14 participants were involved in generating the results from a number of different musical backgrounds.

Section 3.5 demonstrates the following points:

3. Timbre space and timbral relationships are not only an engineering model, but also a psychoacoustic one.

4. Knowledge of timbral relationships is partly independent of musical training, and is a natural part of perception.
5. Relationships between timbres are perceived in a logical, structured manner.
6. Perception of timbre is continuous, rather than categorical.
7. Instruments' timbre spaces are not neatly separated in perception, but have intersecting characteristics.
8. A mental timbre space describing a range of timbral form, as imparted from the sonic information in a set of template sounds, can be successfully used to match the features of stimuli which are not part of the template. This, then, matches a wide range of features and avoids the effects of pitch, loudness and duration on the timbre comparisons.

6.2.3 Chapter 4 : Analysis-Synthesis Model

1. The time-varying frequency spectrum is a greatly used, convenient and effective class of models for investigation of the properties of sounds (Section 4.2).
2. A large number of spectral techniques have been developed, and many techniques are available. Discrete Fourier Transforms and similar techniques are most often used in the literature (Section 4.3).
3. The analysis-synthesis system in this research is used for generating time-varying frequency-domain distributed representations of the 153 input stimuli which may be used for the extraction and analysis of spectral features as detailed in Chapter 5.
4. The chosen technique is based on a multiple time/frequency resolution analysis, which adapts to the conditions of the sound based on a set of periodicity metrics. This was found to be a concept whose implementation could be an improvement over single-level analysis, but which was found to be difficult to implement.

6.2.4 Chapter 5 : Timbral Feature Extraction and Analysis

1. Acoustical features are extracted in this research to enable analysis of the relationships between acoustical forms and perceptual structure of timbral difference.

These features can be extracted from the spectral representations produced by the system described in Chapter 4 for the 153 input stimuli. The feature specifications are developed from those previously found to be important in Sections 2.9 and 2.10 (Section 5.2).

2. The feature extraction and analysis method is concerned with finding specific aspects of acoustical difference which relate to perceptual differences, rather than general areas of importance. The features are either static measures or time-varying quantities. The chapter considers hierarchical decomposition as an aid to understanding perceptual structure. Objective algorithmic extraction and analysis techniques are used (Section 5.3).
3. The features cover the areas of static measures of amplitude envelope form, time-varying measures of strong partial characteristics, and time-varying measures of spectral shape. They do not refer specifically to formant structures, sustain/release portions, spectral element widths/noise bands, or synchrony of partials (Section 5.3).
4. The feature extraction algorithms are outlined in Section 5.4.
5. Feature analysis is concerned with finding which of those in the extracted set contribute most to the perceived differences between groups of stimuli. Time-varying features are characterised by a number of statistical metrics to achieve static values which can be compared. This results in a total of 335 features for all the 153 sounds (Section 5.5).
6. The dimensionality of a particular timbral description is determined by the best set of axes, in some sense, for distinguishing between groups of stimuli at all levels of detail (of interest) in a perceptually structured manner (Section 5.5).
7. The algorithm used for finding which subset of the features is appropriate for distinguishing particular groups of stimuli (AXESDIST) is described in Appendix B.
8. The first part of the feature set analysis correlates the scaled perceptual dimensions developed in Chapter 3 with the acoustical features. Only the first dimension in the 2D and 3D solutions provides a high correlation value with the feature set, indicating a more complex hidden structure which requires multiple simple dimensions to explain the details (Subsection 5.6.1).

9. The second part of the analysis finds those features which best distinguish particular groups of sounds using the AXESDIST program. This includes general high level, instrument group, and lower level distinctions. A considerable range of feature types are found to be relevant, and so the overall dimensionality can be considered quite large (Subsection 5.6.2).
10. The third part of the analysis considers those aspects that affect the number of dimensions/features necessary to describe timbral differences and also concludes that the dimensionality of timbre space is likely to be much higher than 3-5 degrees of freedom (Subsection 5.6.3).

Overall, the experiments of Section 5.6 demonstrate the following points:

11. Features can be extracted from the spectral form which enable sound qualities to be distinguished; and those features are not arbitrarily linked to perceptual differences, but show logical relationships.
12. Hierarchical discrimination of sound qualities is an effective system for structured decomposition of timbre space.
13. The dimensionality of timbre space is considerable and the perceived differences between all perceptible stimuli cannot be completely explained by a low number of dimensions.

6.3 Novel Aspects

6.3.1 Chapter 2 : Perspectives and Research into Timbre

1. The chapter is the most comprehensive overview of the sound timbre literature and concepts to date.
2. The analysis of the components defining the concept of timbre has not been previously attempted.
3. The in-depth consideration of the definition, factors affecting, dimensionality and structure of timbre space is more wide-ranging than previously considered.

4. The review of the spectral aspects and features relating to timbre in the literature is the most comprehensive to date.

6.3.2 Chapter 3 : Perceptual Study

1. This chapter contains an overview of the processes of perception applied to the study of timbre, of which no other to this depth is known to the author.
2. The perceptual study uses the largest number and range of stimuli to date of a timbre perception experiment, and thus can claim to be more comprehensive and universally applicable than previous attempts studying the same sorts of sounds.
3. The study uses the largest range of statistical metrics of any timbre perception experiment to date in order to reinforce the conclusions from several perspectives.
4. The study successfully pioneers the use of a template matching technique in judging timbral similarity to avoid the effects of non-timbral characteristics impinging on the judgements.
5. The study draws the most wide ranging conclusions to date (as described previously).

6.3.3 Chapter 4 : Analysis-Synthesis Model

1. The system works with a wide range of sounds without user intervention but has adaptive time-frequency trade-off linked to the periodic/stability conditions within the stimuli without a knowledge-based system and with a frequency-domain result. The implementation of this particular combination of features is novel.

6.3.4 Chapter 5 : Timbral Feature Extraction and Analysis

1. The experiments use a considerably greater number, range and detail in the stimulus set than has been previously used, and can thus claim to be more comprehensive and universally applicable than previous work.
2. The feature extraction and analysis study is based on a much larger acoustical feature set than has been previously examined.

3. The feature searching algorithm (AXESDIST) uses an especially designed exhaustive enumeration technique to examine the feature set in a more appropriate manner than is possible with the standard scaling and discrimination techniques used by previous researchers.
4. This chapter includes an empirical examination of dimensionality and hierarchical decomposition of acoustical timbre space with relation to perception, which has not been attempted before.
5. The study is the most comprehensive single examination of the relationship between the spectral/acoustical form and timbre perception to date through its use of a number of related experiments.

6.4 Potential Further Work

Other researchers might find it profitable to follow up the work detailed in this thesis in the following areas:

6.4.1 Chapter 4 : Analysis-Synthesis Model

1. Other methods of creating an adapting analysis could be considered. For example, different adapting metrics, relationships between levels of detail, and derivation of the composite form using those metrics.
2. Real-time analysis would have a number of advantages, allowing not only efficient configuration, but also application in many other areas of music and aural technology apart from converting samples to a form suitable for further investigation of acoustic/timbral relationships, as in this research.
3. The distortions imparted by the peripheral hearing system could be considered more deeply.

6.4.2 Chapter 5 : Timbral Feature Extraction and Analysis

1. To improve universality of any results obtained, future work could take an even bigger viewpoint on the problem of timbre research; considering more stimulus types

and a wider range of spectral features. Although the necessary computing power for such investigations is becoming available to all, the fundamental problems of algorithmic effectiveness and the interpretation of results will not go away, however.

2. The comparison of different algorithmic techniques for feature extraction could be beneficial to researchers in determining the most effective ways to progress. Even such basic attributes of sounds as the fundamental frequency have no recognised “best” method of extraction.
3. A deeper consideration of the time-varying nature of the parameters of sound samples other than the amplitude envelope could be profitable. For a long time it has been known that spectral parameter evolution is important, but the nature of evolution is a not well understood.
4. Particular features which could be useful to consider in future are formant structures, sustain/release portions, spectral element widths, and synchrony of partials. Also, the relationship between spectral features is of potential interest. That is, not only extracting the raw features, but also considering particular known groupings (lower harmonics, odd/even harmonics, distribution of noise bands and so on).
5. More complex statistical algorithms than AXESDIST could be developed to find other data patterns.
6. Experimentation is required to obtain greater understanding of the effects of context, perceived dimensionality and the structure of timbral perception.
7. The manipulation of spectral axes could be of interest. Such an ability would allow researchers to test the interpretations of analysed results in a direct manner.
8. The ultimate aim of these sorts of investigation is to find the fundamental axes of timbre space. That is, the acoustical forms which describe all perceptible sound qualities in a structure which mimics the processes of the auditory cortex.

6.5 Confirmation of Hypothesis

Timbre space is a perceptually-structured complex multi-dimensional form, which is ineffectively and imprecisely characterised by a low number of exclusive single dimensions or general descriptions, but can be represented by grouped features derived from time and frequency domain representations.

It is intuitively true that structure exists in the perception of timbral form, but it is shown scientifically by the research detailed in Chapter 3. That work shows that structured timbre perception exists in a similar form with all the participants, whatever their musical background. Such a conclusion is a positive step in understanding that there are logical relationships to be found in studies of timbre perception. The author believes that, although the analysis only considered the results for 14 participants and 153 sounds, that future studies considering more participants and stimuli will find similar structures, which could be developed into a comprehensive model of timbral processes. Such a model, however, must account for the currently little-understood human ability to achieve improved timbral discrimination through exposure/learning, and so have greater timbral appreciation than is naturally present in all persons. That is, perception of timbre is structured and similar between people, but may yet be found to have complex experience-dependent properties. Those aspects were not explored by this research.

Timbre perception is complex and multi-dimensional. This has been consistently shown in previous studies from both perceptual and acoustical viewpoints (Chapter 2) as well as in this research. The degree of the dimensionality has not been established to date. That the majority of timbral variation among a small number of orchestral sounds can be described by a low number of axes (3 to 5) is not proof of a dimensional form for all perceptible timbral differences. That this research has considerably more stimuli over a wider perceptual range than many previous studies (Section 2.5) lends weight to the conclusions of Chapter 5 that more dimensions than 3-5 are required to precisely explain all the perceptual variation present. If the stimulus set included sounds from many other sources such as everyday sounds and in greater number, it seems likely that more dimensions would result. At some point, redundancy would probably curtail further increases, but the overall conclusion is still that timbre is ineffectively and imprecisely characterised by a low number of dimensions.

The dimensions which have been investigated in Chapter 5 are coupled in many ways. It might be possible to extract orthogonal dimensions from timbral model representations. However, the evidence in this research and previous studies suggests that timbre is not a phenomenon that lends itself to description through exclusive single dimensions. Neither are the other dimensions of sound (pitch, loudness, duration and environment) decoupled from timbre. It would be a great advantage to areas such as those listed in Section 1.2 in which timbre plays a role, were the perceptual and acoustical axes of timbre a low number of universal uncoupled axes. At this stage, such a prospect seems unlikely.

It is apparent from the dissertation in Chapter 2 that it is almost impossible to summarise what timbre means in a concise manner to the satisfaction of all researchers in the field. The complexity and lack of concrete information concerning timbre has led to a large number of different descriptions and timbre spaces over which research work can be considered valid. This means that general descriptions of a few lines fail to provide a complete, effective and precise assessment of what timbre means. The unambiguous method of definition is to outline the timbre space associated with the research. The alternative is to define timbre in as broad terms as possible, which is likely to be very uninformative.

Time and frequency domain representations remain the models of choice in research into sound qualities for the reasons outlined in Section 4.2. The growing body of information is gradually building a more solid base for understanding timbre. It would be wrong to neglect other forms, but the author concludes that time and frequency representations are adequate for the purpose of investigating timbre. The structure that is built to support investigation can have many forms, due to the lack of knowledge concerning perceptually-appropriate systems of organising timbre. The structure used in Chapter 5 used a hierarchical form, whereby grouped features described regions of timbral variation within a part of the tree. That this proved a systematic and useful scheme does not imply that it is totally perceptually accurate, but the results show that it is an effective model.

6.6 Concluding Remarks

“In the year 1800 Volta assembled a large battery of the electric cells that he had recently invented, and he connected the total array to a pair of metal rods inserted in his ears. Then he closed the switch. He felt a jolt in the head, he tells us, followed by a noise like the boiling of thick soup. It seems that he decided not to repeat the experiment.” [198]

Volta’s early experiment concerning the operation of the ear and its resulting timbral effect is an example of how past research can appear crude and limited in effect from the modern perspective. From a low knowledge base and through the limitations of equipment, researchers’ progress is restricted. People such as Helmholtz ([83]), Risset and Mathews ([172]), Grey ([71]) and others provided considerable insight into timbral form commensurate with their equipment, the knowledge base from which they worked and their (considerable) intellects. It is apparent that research into complex phenomena such as the human hearing system only progresses through constant re-evaluation of the facts concerning the different components of the field. A single set of experiments is only a step toward a new way of understanding, it will not revolutionise thinking on its own ([31]). This is particularly true in an topic as complex as sound timbre, where the algorithms and assumptions made in the research may influence the results significantly.

“We seek knowledge about controlling timbre synthesis that is both *quantitative* and *universal*” [62]

A major contribution of this research work to the field of sound timbre is to postulate enhanced explanations as to what occurs in the perception of sound quality. The author believes that the next stage should involve the use of many more sound instances than have been considered here (in the thousands, from a much more diverse range of timbral form); with a number of perceptual testing schemes with tens of participants; and analysis, synthesis and modification schemes which test a vast range of aspects of acoustic forms. Such scaling of experimental technique is required to understand the universal structure of timbre perception and its links to acoustical forms. If such a scale of approach is unreasonable, then at least it is important that researchers begin to establish a set of

conditions (timbre space) which facilitate different persons working on the same problem producing results which can be compared.

A particular problem is that previous research has often lost sight of the “purpose” of timbral perception. If the aim of timbral analysis in the ear is only to distinguish sound sources (such as instruments) in a broad way, then there is a lot of redundant computation going on. Aural perception in humans is very good at distinguishing timbral nuances, as it has evolved in a way to permit complex aural analysis in a wide range of situations.

Therefore, research should be aiming to understand all perceptible timbral variation, not just the first 80% in a set of a few orchestral sounds. Often that research has considered that the majority, such as 80%, of the variation is the whole story, when the nuances (and the associated higher dimensionality) are hidden in the last 20%.

“... no single method can uncover the truth, only a part of the truth. Different aspects of the truth, then, can only be uncovered by a convergent system of methods.” [76]

A significant restriction to understanding timbre through analysis of acoustical form has been a lack of computing power. In the near future, however, desktop computing equipment will be sufficiently powerful to consider timbre on the sort of scale described above. Real-time analysis, modification and synthesis of complex timbral forms would also be of great benefit, by permitting interactive experimentation.

“Sounds are intrinsically complex. Musical instruments have a complex physical behaviour ... human musicians ... introduce intricacies both intentionally and unintentionally.” [173]

The nature and form of sound timbre perception has been regarded as, at one extreme, impossible to comprehend and evaluate, and at the other, a low dimensional construct that basically corresponds to simple properties of the acoustics of instruments. As is often the case, the truth is probably somewhere in between. It may be that the dimensionality of that which can be heard is similar to the size of the periodic table in chemistry, where current research is still in the equivalent age of earth, wind, fire and water. Or the answer may be as elegant and compact as Newton’s laws of motion. Whatever the case, there is

always a temptation to be too simplistic rather than complex enough, in order to profess universal truths from the results of single experiments.

A problem in understanding timbre is coping with the level of detail under consideration. Taking regard of very small nuances can miss the big picture (for example, two identically-played notes on the same instrument, which will always be very slightly different physically). Yet, sometimes some smaller effects can be masked by the larger picture, so that in perceptual tests people do not take account of all parts of the sound. Computer analysis is objective in that it can “see” minor details, when a human might miss out on that aspect through familiarity or whatever, but that it also the computer’s downfall, in that it takes regard of information that is of no interest.

As consistently mentioned in this thesis, results from timbre experiments are applicable within an associated timbre space. Such results could apply to other situations, but without further experimentation from a variety of perspectives such extensibility remains unproven. Having more information should lead to the ability to achieve more universal understanding. Yet the processes of dealing with vast amounts of data must be organised enough to achieve progress. Greater information can lead to greater incomprehension, not the opposite. There is also always the possibility of being so general that the nuances of sound timbre which furnish much about the perception of sounds are missed in a desire to comprehend the overall picture.

“Composers would like to work by defining and specifying perceptual parameters within a given sound, rather than synthesis parameters” [146]

It is not only desirable to analyse sound timbre, but also to reverse the process and precisely manipulate the character of sounds. For example, to be able to create timbral objects, which represent tonal change and can be modified, combined, scaled and translated to different parts of timbre space to apply specific modifications to different sounds ([20], [35]). There is the potential for access to an unlimited World of timbres ([173]). But currently composers must tweak and stumble across useful sounds in the process of looking for timbral forms. While this is a valid method of searching for inspiration, to convey a timbral image in the mind of the composer requires an ability to achieve that image in practice through specification. Beethoven when deaf was still able to convey his timbral images as he knew how to control the instruments available to him and

could specify his intentions. The equivalent of the sort of effects that can be achieved with modern computer graphical techniques should be possible in sound timbre. But the visual effects can often be specified exactly with many different images, and the sonic effects can only be achieved through trial and error and in certain situations.

Another way of looking at the timbre control problem is to consider it from an instrumental perspective. In specifying a timbral score the composer might desire enough control to be able to achieve a particular sound quality. That control needs to be logical and learnable, but not necessarily discrete and orthogonal. Human “tools” do not have discrete axes of control, as people can cope with parallel aspects and complex interactions between many dimensions. This also extends to musical instruments. But, the properties of instruments can be practised until a high level of control over timbre is achieved. The problem is that if the dimensionality of the entire space of timbral difference is as high as the author suggests, understanding and controlling all of timbre space is not like a single instrument which can be practised until mastered, but a much more complex interactive field of control.

APPENDIX A

Descriptions of Sound Samples

This appendix describes the sound samples used as input for the systems described in this thesis. They were selected by the criteria given in Chapter 2.

The major group of sounds (numbers 1-110) is taken from the McGill University Master Samples (MUMS) archive. The descriptions given are those of the MUMS accompanying notes. It seems apparent that these are not always complete. For example, some nuances such as vibrato are not always noted when present. With the MUMS sounds, the pitches were chosen to balance the two requirements of a note considerably within the range of the instrument (to achieve a characteristic sound), while also attempting to equalise pitches between instruments (at least to octaves of “A”) to reduce any additional complications that might arise in analysis or psychological testing.

The next group (numbers 111-130) are synthesised sounds from an Orchid NuSound PnP IBM-PC sound card’s wavetable ROM. These represent more “synthetic” types of sounds

than the MUMS group, which are samples of real instruments. The group 131-146 are simple test tones, and 147-153 are miscellaneous samples included for testing whether some unusual sounds are as valid as more “instrumental” sounds in the system under consideration.

The sounds are very slightly coloured by noise which impinged on their translation to the final sound file style of 22050Hz sample rate in 16 bit Mono Microsoft WAV format. They were also trimmed at the very ends to remove low level noise, and the peak amplitude adjusted to make them comparable.

In the following table, MUMS codes refer to (CD number) / (track number)-(note number).

Table A.1: Descriptions of Sound Samples in Use

| Code | MUMS Code | Description | Length(s) |
|-------------|------------------|------------------------------------|------------------|
| 1 | 1/01-16 | violin, bowed, vibrato A4, stopped | 3.53 |
| 2 | 1/01-17 | violin, bowed, A4, open | 3.13 |
| 3 | 1/02-16 | violin, muted, vibrato A4, stopped | 3.28 |
| 4 | 1/02-17 | violin, muted, A4, open | 3.30 |
| 5 | 1/03-16 | violin, pizzicato A4, stopped | 0.33 |
| 6 | 1/03-17 | violin, pizzicato A4, open | 0.53 |
| 7 | 1/05-16 | violin, martele, A4, stopped | 0.25 |
| 8 | 1/05-17 | violin, martele, A4, open | 0.45 |
| 9 | 1/06-24 | viola, bowed, vibrato A4, stopped | 3.39 |
| 10 | 1/06-25 | viola, bowed, A4, open | 3.12 |
| 11 | 1/07-22 | viola, muted, A4, open | 1.48 |
| 12 | 1/08-22 | viola, pizzicato A4, open | 0.64 |
| 13 | 1/10-22 | viola, martele A4, stopped | 0.39 |
| 14 | 1/11-24 | cello, bowed, vibrato A3, stopped | 4.51 |
| 15 | 1/11-25 | cello, bowed, A3, open | 3.08 |
| 16 | 1/12-22 | cello, muted, A3, open | 4.70 |
| 17 | 1/13-22 | cello, pizzicato A3, open | 1.86 |
| 18 | 1/15-23 | cello, martele A3, stopped | 1.42 |

CONTINUED →

Table A.1: Descriptions of Sound Samples in Use

| Code | MUMS Code | Description | Length(s) |
|------|-----------|-------------------------------------|-----------|
| 19 | 1/16-10 | double bass, bowed, A1, stopped | 2.49 |
| 20 | 1/16-11 | double bass, bowed, A1, open | 2.19 |
| 21 | 1/17-10 | double bass, muted, A1, stopped | 1.81 |
| 22 | 1/17-11 | double bass, muted, A1, open | 2.85 |
| 23 | 1/18-10 | double bass, pizzicato, A1, stopped | 2.37 |
| 24 | 1/20-10 | double bass, martele, A1, stopped | 0.98 |
| 25 | 8/27-01 | acoustic bass, plucked A1 stopped | 2.43 |
| 26 | 8/38-02 | acoustic bass, harmonics A2 | 3.58 |
| 27 | 5/16-01 | deep electric bass, pop style A3 | 4.17 |
| 28 | 2/01-10 | flute, vibrato A4 | 3.91 |
| 29 | 2/02-22 | flute, flutter A5 | 4.88 |
| 30 | 2/03-08 | piccolo A5 | 2.97 |
| 31 | 2/04-20 | piccolo, flutter A6 | 3.51 |
| 32 | 2/05-15 | alto flute, vibrato A4 | 3.40 |
| 33 | 2/06-22 | bass flute, vibrato A4 | 4.09 |
| 34 | 2/07-10 | bass flute, flutter A3 | 2.83 |
| 35 | 2/08-12 | oboe, A4 | 2.42 |
| 36 | 2/09-06 | English horn, A3 | 3.00 |
| 37 | 2/10-08 | B flat clarinet, A3 | 3.89 |
| 38 | 2/11-03 | E flat clarinet, A3 | 4.47 |
| 39 | 2/12-09 | bass clarinet, A2 | 2.42 |
| 40 | 2/13-04 | contrabass clarinet, A1 | 3.35 |
| 41 | 2/14-12 | bassoon, A2 | 4.80 |
| 42 | 2/15-12 | contrabassoon, A1 | 3.13 |
| 43 | 3/13-02 | bass saxophone, A1 | 2.52 |
| 44 | 3/14-10 | baritone saxophone, A2 | 4.37 |
| 45 | 3/15-10 | tenor saxophone, A3 | 2.47 |
| 46 | 8/67-05 | tenor saxophone, growl A3 | 3.86 |
| 47 | 8/75-07 | tenor saxophone, subtones A2 | 4.10 |

CONTINUED →

Table A.1: Descriptions of Sound Samples in Use

| Code | MUMS Code | Description | Length(s) |
|-------------|------------------|---------------------------------------|------------------|
| 48 | 3/16-09 | alto saxophone, A4 | 2.92 |
| 49 | 8/80-01 | alto saxophone, growl A4 | 4.29 |
| 50 | 8/82-03 | alto saxophone, scream A5 | 0.80 |
| 51 | 3/17-09 | soprano saxophone, A5 | 3.44 |
| 52 | 2/16-28 | C trumpet, A5 | 4.28 |
| 53 | 2/17-28 | C trumpet, harmonics with stem out A5 | 3.45 |
| 54 | 2/18-11 | Bach trumpet, A4 | 3.89 |
| 55 | 7/15-01 | B flat trumpet, hard attack A3 | 2.30 |
| 56 | 7/20-03 | B flat trumpet, bucket mute, loud A3 | 2.66 |
| 57 | 8/49-03 | B flat trumpet, bucket mute, soft A3 | 2.23 |
| 58 | 8/54-05 | B flat trumpet, cup mute, loud A3 | 3.11 |
| 59 | 8/60-05 | cornet with straight mute, A3 | 2.70 |
| 60 | 2/19-20 | French horn, A3 | 1.91 |
| 61 | 2/20-20 | French horn, muted A3 | 1.72 |
| 62 | 2/21-05 | alto trombone, A4 | 2.06 |
| 63 | 2/22-06 | tenor trombone, A2 | 1.27 |
| 64 | 2/23-06 | tenor trombone, muted A2 | 1.99 |
| 65 | 2/24-05 | bass trombone, A1 | 1.34 |
| 66 | 2/25-09 | tuba, A2 | 4.39 |
| 67 | 2/26-02 | trombone pedal note, A1 | 2.47 |
| 68 | 3/02-01 | 9' Hamburg Steinway, loud A0 | 3.27 |
| 69 | 3/02-13 | 9' Hamburg Steinway, loud A1 | 3.48 |
| 70 | 3/02-25 | 9' Hamburg Steinway, loud A2 | 3.07 |
| 71 | 3/02-37 | 9' Hamburg Steinway, loud A3 | 3.62 |
| 72 | 3/02-49 | 9' Hamburg Steinway, loud A4 | 3.02 |
| 73 | 3/02-61 | 9' Hamburg Steinway, loud A5 | 2.98 |
| 74 | 3/02-73 | 9' Hamburg Steinway, loud A6 | 2.64 |
| 75 | 3/02-85 | 9' Hamburg Steinway, loud A7 | 1.99 |
| 76 | 3/04-29 | symphonic marimba A4 | 0.96 |

CONTINUED →

Table A.1: Descriptions of Sound Samples in Use

| Code | MUMS Code | Description | Length(s) |
|------|-----------|--|-----------|
| 77 | 3/05-05 | xylophone A4 | 1.09 |
| 78 | 3/06-17 | vibraphone, hard mallet A4 | 1.65 |
| 79 | 8/86-05 | vibraphone, soft mallet A3 | 3.27 |
| 80 | 3/07-17 | vibraphone, bowed A4 | 1.83 |
| 81 | 3/08-03 | glockenspiel, brass beater A5 | 1.59 |
| 82 | 3/09-10 | crotales, brass beater A6 | 2.54 |
| 83 | 3/10-10 | tubular bells, A4 | 2.38 |
| 84 | 3/11-07 | snare drum, hit | 0.23 |
| 85 | 3/11-20 | 10" tom-tom | 0.43 |
| 86 | 3/11-25 | small timbales | 0.96 |
| 87 | 3/11-32 | conga, open tone | 0.38 |
| 88 | 3/11-33 | conga, closed tone | 0.37 |
| 89 | 3/11-35 | conga, slide | 0.89 |
| 90 | 3/11-36 | tumba, open tone | 0.54 |
| 91 | 3/11-38 | orchestral bass drum | 1.15 |
| 92 | 3/11-39 | rock bass drum | 0.15 |
| 93 | 3/11-40 | tympani, C2 | 2.08 |
| 94 | 3/12-01 | 20" Chinese cymbals, crash | 2.29 |
| 95 | 3/12-03 | 15" Turkish cymbals crash with soft mallet | 3.86 |
| 96 | 3/12-04 | orchestral cymbals, crash | 3.30 |
| 97 | 3/12-06 | small gong | 4.86 |
| 98 | 3/12-12 | alpenglocken | 1.04 |
| 99 | 3/12-14 | cencerros | 1.52 |
| 100 | 3/12-18 | agogo bells | 0.82 |
| 101 | 3/12-20 | finger cymbals | 1.43 |
| 102 | 3/12-22 | sleigh bells | 1.17 |
| 103 | 3/12-26 | large triangle | 3.95 |
| 104 | 3/12-31 | temple blocks | 0.22 |
| 105 | 3/12-42 | cabasa, roll | 0.77 |

CONTINUED →

Table A.1: Descriptions of Sound Samples in Use

| Code | MUMS Code | Description | Length(s) |
|------|-----------|-------------------------------|-----------|
| 106 | 3/12-43 | ratchet | 1.00 |
| 107 | 3/12-45 | tambourine, pop | 0.75 |
| 108 | 3/12-49 | cuica | 1.24 |
| 109 | 3/12-60 | Chinese ascending gong | 1.96 |
| 110 | 5/56-04 | synthesized analogue bass, A1 | 3.60 |
| 111 | - | accordion A2 | 0.91 |
| 112 | - | bass guitar A1 | 0.94 |
| 113 | - | celesta A1 | 0.93 |
| 114 | - | church organ A2 | 1.00 |
| 115 | - | contra bass A0 | 0.99 |
| 116 | - | dulcima A1 | 0.99 |
| 117 | - | glockenspiel A1 | 1.00 |
| 118 | - | halo pad A2 | 1.00 |
| 119 | - | Hammond organ A2 | 0.86 |
| 120 | - | harmonica A1 | 0.96 |
| 121 | - | metallic pad A1 | 1.00 |
| 122 | - | music box A1 | 1.00 |
| 123 | - | percussive organ A1 | 0.84 |
| 124 | - | rock organ A1 | 0.90 |
| 125 | - | slap bass A1 | 1.00 |
| 126 | - | shantai A2 | 0.95 |
| 127 | - | steel drum A2 | 0.99 |
| 128 | - | synth bass A0 | 0.82 |
| 129 | - | tubular bell A1 | 0.99 |
| 130 | - | vibraphone A1 | 1.00 |
| 131 | - | sine wave A0 | 1.00 |
| 132 | - | sine wave A1 | 1.00 |
| 133 | - | sine wave A2 | 1.00 |
| 134 | - | sine wave A3 | 1.00 |

CONTINUED →

Table A.1: Descriptions of Sound Samples in Use

| Code | MUMS Code | Description | Length(s) |
|-------------|------------------|---|------------------|
| 135 | - | sine wave A4 | 1.00 |
| 136 | - | sine wave A5 | 1.00 |
| 137 | - | sine wave A6 | 1.00 |
| 138 | - | sine wave A7 | 1.00 |
| 139 | - | sine sweep (low→high) | 1.00 |
| 140 | - | square A3 | 1.00 |
| 141 | - | 50% pulse A3 | 1.00 |
| 142 | - | sawtooth A3 | 1.00 |
| 143 | - | pink noise | 1.00 |
| 144 | - | pink noise mixed with sine A7 | 1.00 |
| 145 | - | white noise | 1.00 |
| 146 | - | white/sine A1/white | 1.00 |
| 147 | - | warning alarm C3 | 1.00 |
| 148 | - | dog bark | 0.50 |
| 149 | - | klaxon | 0.97 |
| 150 | - | lion roar | 1.00 |
| 151 | - | muted trumpet progressive growl shift F#3 | 0.92 |
| 152 | - | breaking glass | 0.99 |
| 153 | - | American train whistle | 1.00 |

APPENDIX B

Mathematics Relevant to This Research

This appendix describes the mathematical techniques cited and used within this thesis.

Mathematics is used in this research both to:

1. Confirm results of other authors and the expectations of this author, and
2. Explore the result spaces of the experiments to form new conclusions.

The use of the wrong techniques to analyse the result data can cause incorrect confirmation of expectations and incorrect conclusions to be drawn, but use of the right technique can uncover hidden information in complex data structures. Throughout the mathematical analyses in this thesis, a number of related techniques are used in order to achieve a balanced view of the data forms. It is not necessary to understand the fine

details of the most complex techniques such as principal components analysis ([119]) in order to use them to generate meaningful results. It is necessary to understand their background and limitations, however. This Appendix briefly describes the differences between the data analysis techniques cited or used in this thesis, and also some of the associated methods. In so doing it shows why certain techniques have been used in this thesis and what other options exist.

The simpler metrics/techniques listed below are described in mathematical terms. For the more complex techniques, the reader is advised to study the texts specific to the topics which are cited. Unless stated, all techniques are available in SPSS version 6.1 and the reader is also advised to read the texts associated with that package, particularly [140] (hypothesis testing, correlation/regression, Kolmogorov-Smirnov tests) and [141] (discriminant analysis, factor analysis, principal components analysis, cluster analysis, distance metrics, multidimensional scaling). A good general statistics text is [79].

B.1 Basic Statistical Methods

In trying to understand a complex data form, it is often a sensible first step to analyse the general patterning through such metrics as the following:

1. Averages; arithmetic mean, geometric mean, mode, median. These indicate the general tendency of a set of data points. The method chosen necessarily defines what the answer implies. For example, in Subsection 3.5.3 modal responses are used, as the most frequent average makes more sense when describing the average selection from a set of discrete values than an arithmetic mean.

$$\textit{arithmetic } \mu = \frac{1}{N} \sum_{i=1}^N X_i \tag{B.1}$$

$$\textit{geometric } \mu = \left(\prod_{i=1}^N X_i \right)^{\frac{1}{N}} \tag{B.2}$$

$$\textit{modal } \mu = X_{\textit{most frequent value}} \tag{B.3}$$

$$\textit{median } \mu = X_{\textit{central value in value-ordered list}} \tag{B.4}$$

The arithmetic mean is used in characterising time-varying forms in Subsection 5.5.1.

2. Metrics of variation; variance, standard deviation. These are standard indicators of variation around an average value based on the sum of deviations:

$$std.dev. = \sigma_X = \sqrt{variance} = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}} \quad (B.5)$$

Standard deviation is used in characterising time-varying forms in Subsection 5.5.1.

3. Metrics of distribution; percentage classification in different ranges/categories, tests for similarity to standard distributions (Kolmogorov-Smirnov goodness-of-fit), distribution shape (skewness, kurtosis). These more complex metrics aim to describe the general distribution of data. Classification summaries and Kolmogorov-Smirnov tests are used in Subsection 3.5.1.
4. Metrics of slope/position; linear regression, linear correlation . These techniques attempt to find patterns in a line in the data. Regression attempts to find a line of best fit to the data and correlation attempts to find the strength of the linear relationship between two variables. Correlation values range from -1 (perfect opposite relationship) through 0 (no linear relationship) to 1 (perfectly similar relationship). In between, the form is considered weak at a magnitude of 0.2, moderate at 0.5 and strong at 0.8 ([79]). If the data can be considered generally linear, or the strength of any linear relationship is of interest, then these techniques can be used to summarise the general tendency of the data (regression) or the similarity of two sets of data (correlation).

$$Pearson\ corr.\ coeff. = \frac{\sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y)}{(N - 1)\sigma_X\sigma_Y} \quad (B.6)$$

Regression :

$$\begin{aligned} slope = b &= \frac{\sum_{i=1}^N [(X_i - \mu_X)(Y_i - \mu_Y)]}{\sum_{i=1}^N (X_i - \mu_X)^2} \\ intercept = a &= \mu_Y - b\mu_X \\ Y &= a + bX \end{aligned} \quad (B.7)$$

Correlation is used in Subsection 3.5.2 and the regression slope in Subsection 5.5.1.

5. Hypothesis Testing. These techniques are for the purpose of testing the plausibility of a null hypothesis H_0 to determine the likelihood of a research hypothesis H_1 . These tests produce measures which indicate the probability (p value) of a certain outcome which can be compared with a predetermined cutoff point or significance level (α). The most common of these types of technique are Z, T and ANOVA tests. However, they are not of use in this research work. What is of use is the Kolmogorov-Smirnov tests mentioned previously used for determining fit to standard distributions (Subsection 3.5.1). The p values resulting from these indicate significance for greater values.

B.2 Multivariate Scaling Algorithms

Techniques for finding representative variable sets where the number of input variables is greater than the number which might reasonably represent the majority of variation among the input variables have seen considerable use in timbre research. This is due to the nature of timbre being highly multidimensional and a desire on the part of researchers to find fundamental axes of variation which can explain perceived differences in a compact, yet clear manner. That is, not merely a low dimensional representation (of which the time waveform is comprehensive), but rather a dimensional form with redundant information removed, yet information distributed in an interpretable fashion.

Normally, researchers are faced with a considerable number of dimensions in which the useful information is spread over the variable set and mixed with low level information which is possibly not of interest. An example of this is the data presented in Chapter 3. It is often the case, as well, that it is desirable to visualise the information on a low dimensional plot, even if some accuracy is lost, in order to better understand the data set. The techniques necessarily require more correlation for greater reductions in the number of dimensions to minimise the loss of accuracy.

1. Principal Components Analysis (PCA, see [119] for more extensive details).

The aims of PCA are to:

- (a) Find relatively independent variables.
- (b) Reduce large sets of variables to smaller, more meaningful/interpretable sets.

- (c) Determine the necessary number of dimensions to represent the set with acceptable loss of information.
- (d) Test hypotheses about the structure/relationships of variables.

The effect of PCA is to take p variables $X_1 \dots X_p$ and find combinations of these to produce indices (principal components/factors) $Z_1 \dots Z_p$ that are uncorrelated and are each a linear combination of the variables ($Z_i = a_{i1}X_1 + \dots + a_{ip}X_p$ where $a_{i1}^2 + \dots + a_{ip}^2 = 1$). The indices are, also, ordered such that Z_1 displays the largest amount of variation and Z_p the least. The idea is, then, that the Z_i s for the lower i values can be kept and the others discarded as, in theory, the vast majority of variation among the original p is now represented in fewer variables (the underlying dimensions of the data). As mentioned previously, there must be significant linear correlation among the input variables to achieve this reduction. Studying the variance explained by the individual PCs can aid in understanding how successful the technique has been. The technique is:

- (a) Code the input variables to have zero means and unit variances.
- (b) Find the covariance matrix.
- (c) Find the eigenvalues (variances of PCs) and eigenvectors (coefficients of each PC, a_{ij}).

The technique is sensitive to:

- (a) Outliers.
- (b) Poor correlations between variables.
- (c) Very small sample sizes.

The technique is used in Subsections 3.5.4 and 3.5.5. PCA has also been used or described with respect to timbre studies in [109], [177], [178], [157], [217], [13], [62] and [130].

2. Factor Analysis (FA, see [119]).

The aims, general effect and sensitivities of FA are the same as PCA, but the techniques used are more complicated and exist in greater numbers. These represent alternatives to PCA that could have been used in Subsections 3.5.4 and 3.5.5 had there been more time. The FA model is of the form $X_i = a_{i1}F_1 + \dots + a_{im}F_m + e_i$.

That is, the original variables are described in terms of a weighted sum of factors plus a unique factor e (specificity), uncorrelated with any of the common factors F , specific to the i th equation.

There are an infinite number of ways of establishing the factor loadings (a_{ij}). In fact, PCA can be achieved with FA by restricting the technique. The variations in technique are based on different methods of maximising the quality of fit (the similarity of the observed data and data reproduced from the resulting factor solution). The types of factor analysis available in SPSS are unweighted least squares, generalised least-squares, maximum likelihood, principal axis factoring, alpha method and image factoring. Factor analysis has been used or described with respect to timbre studies in [196], [197], [139] and [99].

3. Multidimensional Scaling (MDS, see [119]).

MDS is also similar to PCA and FA in its aims, but works with proximity ((dis)similarity) data to construct a geometric representation in fewer dimensions than there were originally. This is particularly useful for direct similarity judgement experiments where there is a direct measurement of every case against every other. Where cases are measured individually on a number of variables, it is advisable to use FA/PCA rather than deliberately derive the distance information, although that is possible in SPSS. The advantages of MDS are that the solutions can be more interpretable in their raw form (being based on distance rather than vectors), do not assume linear relationships (as FA/PCA do) and the technique can be metric or non-metric (based on rank) in SPSS. A number of techniques exist under the heading MDS. MDS has been used or described with respect to timbre studies in [107], [218], [131], [216], [71], [19], [87], [72], [70], [74], [159], [93], [130], [123], [43] and [102].

4. Optimal Scaling (OS).

This is a set of procedures provided by SPSS which has similarities with those already discussed. It works with nominal (categorical) rather than interval data, and so could be used with the perceptual study data of Chapter 3. OS is a scaling technique like a nominal, limited version of FA.

5. Other Considerations

The above procedures do not necessarily produce solutions where the resulting axes have a simple relationship to the original variables. Also, if the original data is from

a non-physical form such as the perceptual data of Chapter 3, the resulting axes do not necessarily relate directly to particular acoustical quantities. If the relationships between a scaled perceptual data solution and acoustical quantities is required, then it is necessary to correlate the resulting structure with feature axes. Under SPSS, solutions from PCA and FA can be rotated in order to redistribute the data such that only some of the variables load onto each factor. This should make the structure more easy to interpret under some circumstances.

B.3 Group Discrimination Algorithms

The techniques of Section B.2 attempt to reduce the number of dimensions to a more fundamental set to improve interpretability, understanding of structure and so forth. The techniques in this section attempt to either find the groups within the data, or find which variables result in particular groupings.

1. Cluster Analysis (CA, see [119]).

Cluster Analysis attempts to find groups within the data based on proximity by linking cases together. SPSS provides a range of linking methods (between-groups, within-groups, nearest neighbour, furthest neighbour, centroid clustering, median clustering or Ward's method). These techniques necessarily use all the available variables, which makes them less useful for this research, where the desire is to find which particular acoustical features are relevant in creating particular groups.

Cluster Analysis has been used or described with respect to timbre studies in [130] and [45].

2. Discriminant Analysis (DA, see [119]).

Discriminant Analysis addresses the problem of identifying the variables that are important in distinguishing between groups of cases. A linear combination of independent variables is formed; $D = a_0 + a_1X_1 + \dots + a_pX_p$. This has similarities with the equations used for PCA, except with DA the aim is to achieve a situation where D differs as much as possible between groups by maximising

$\frac{\text{between-groups sum of squares}}{\text{within-groups sum of squares}}$. The values of a_i are calculated to achieve this based on user-chosen groups. Like the other procedures considered so far, this technique

considers all the input variables at the same time. Unfortunately, if there are enough variables (such as the considerable number of acoustical variables in Chapter 5), then there will be enough noise/numerical variation for the algorithm to achieve high quality separation of groups by using large numbers of the variables. That does not, then, aid in understanding which variables are most important, but only shows that given enough data it is possible to discriminate sound qualities.

3. AXESDIST (not available in SPSS).

AXESDIST is a program developed by the author (and used in Chapter 5) which develops the ideas of discriminant analysis in a different way. AXESDIST finds the set of N variables among the input set M which produces the minimum values of $\frac{\text{sum of within-groups distance metrics}}{\text{sum of between-groups distance metrics}}$. That is, the algorithm performs an exhaustive enumeration search of the data space to find the N “best” variables for separating and containing groups of data points. These means that the exhaustive search is determined by

$$\text{loops} = {}^M C_N = \frac{M!}{(M - N)!N!} \quad (\text{B.8})$$

and the number of distances to be calculated within the minimisation quotient. The functional differences to SPSS discriminant analysis are:

- (a) The result is not weighted to achieve the desired groups and so when large numbers of input axes are used, this does not facilitate the algorithm progressively adjusting the variables to fit to the solution, but rather choosing the best unweighted combination for the chosen N .
- (b) A small number of output axes ($N \ll M$) can be specified, which means that the algorithm is trying to choose a best set of N , which is much simpler to understand than the best weighted sum of all M variables.
- (c) AXESDIST has more available control (see below) in the use of distance metrics and the composition of the minimisation than discriminant analysis, which sticks to sums of squares between and within all groups.
- (d) The results are more easily interpreted in terms of the input features involved as they are not “mixed” in the solution.

The algorithm for AXESDIST is as follows:

- Load list of feature data for all features and all stimuli, and scale to make comparable
- Choose groups of stimuli to be used in the search
- Choose a group combination to minimise
- Choose which features must be part of the solution (fixed axes set)
- Choose which features could be part of the solution (variable axes set)
- Choose the size of feature sets to be chosen from the variable axes set
- For all combinations of sets chosen from variable set
 - Calculate the group combination metric based on the current set chosen from the variable set + the fixed set of axes
 - If the metric is in the top 100 lowest results then save to the results list at the appropriate ordered position

The user-specified group combination to minimise is of the form:

$$\frac{\text{sum of within groups distance metrics}}{\text{sum of between groups distance metrics}} = \frac{\text{Within}(nGroup_1, nMethod_1) + \dots + \text{Within}(nGroup_p, nMethod_p)}{\text{Between}(dGroup_{1a} \leftrightarrow dGroup_{1b}, dMethod_1) + \dots + \text{Between}(dGroup_{pa} \leftrightarrow dGroup_{pb}, dMethod_p)} \quad (\text{B.9})$$

This combination equation has components in the numerator which relate to the containment of items within groups. That is, the metrics to minimise are distances within the group of stimuli to be brought together. In the denominator are components which relate to the distinction between groups which are to be maximised, the distances between groups. The user can choose any groups of stimuli from those specified at the second point in the algorithm in the $nGroup_i$ and $dGroup_{j,k}$ positions; they need not be all used nor in a particular order nor the same in numerator and denominator. The *Methods* for containment are:

- (a) Average all points distance.
- (b) Maximum inter-point distance.

The *Methods* for distinction are:

- (a) Average all points distance.
- (b) Minimum inter-point distance.
- (c) Centroid distance.

These methods are simple, to facilitate the use of a large number of data points, complex combination equations, and many combinations of axes in the search without the computation time becoming unreasonable (more than a couple of days per run). The effect is necessarily that the types of groupings/distinctions which can be found are limited to simple clusters in multidimensional space. The reason for being able to specify a set of fixed axes is that if particular axes are required in the solution, or that it is clear from lower dimensional runs that any higher dimensional solutions are likely to contain certain features, then they can be forced into all solutions with the variable sets. This reduces the computational load for higher dimensional solutions by allowing hand-optimisation.

The improvements which could be made in a future algorithm are:

- (a) A wider range of discrimination/containment *Methods* to find more complex shapes in multidimensional space, and more complex group weighting schemes to balance the required discrimination/containment emphasis.
- (b) More automatic dynamic searching, using a technique such as genetic algorithms.
- (c) The use of different distance interpretations apart from the Euclidean space used in this algorithm, by using different Minkowski λ values:

$$\begin{aligned}
 \text{Minkowski distance} &= \|p - q\|_{\lambda} = \left[\sum_{i=1}^N |\xi_i - \eta_i|^{\lambda} \right]^{\frac{1}{\lambda}} \\
 p &= (\xi_1 \dots \xi_N), q = (\eta_1 \dots \eta_N) \\
 (\lambda = 1, \textit{city - block distance}; \lambda = 2, \textit{Euclidean distance}) & \quad \text{(B.10)}
 \end{aligned}$$

After the user has specified the perspective for examining the data, the algorithm takes no further user input. It can thus be considered an objective technique. However, the result is necessarily a particular viewpoint on the data, not *the* answer. AXESDIST is novel in the respect of the use of exhaustive enumeration in exploring the timbre data space, when other researchers have used scaling algorithms. Exhaustive enumeration is a standard mathematical method, but novel in this context.

Glossary

| | |
|---------------------------|---|
| <i>Adaptor</i> | Comparison stimulus (standard) against which other stimuli are judged in a perceptual similarity test. |
| <i>AM</i> | Amplitude Modulation. The process of varying the amplitude of acoustic components to cause a perceptible effect. |
| <i>Amplitude Envelope</i> | Low frequency (less than 20Hz) variation to amplitude of the overall tonal complex. |
| <i>Auditory Cortex</i> | Part of the brain associated with the analysis of aural information captured by the ears. |
| <i>AXESDIST</i> | Statistical procedure devised by the author. Used for finding timbral axes which distinguish between stimulus groups. |
| <i>Bins</i> | Data streams resulting from transformation into the time-varying frequency domain. |
| <i>Critical Bands</i> | Basic “filter bandwidths” of the ear, within which multiple frequency components cannot be cleanly separated in perception and cause beats and roughness. |

| | |
|----------------------------------|--|
| <i>DFT</i> | Discrete Fourier Transform. Transformation from time to frequency domain. |
| <i>Dimensionality</i> | Number of degrees of freedom in the data set. That is, the number of axes used to represent the timbral data achieving correct perceptual ordering of the stimuli. |
| <i>FM</i> | Frequency Modulation. The process of varying the frequency of acoustic components to cause a perceptible effect. |
| <i>Formant</i> | A range of emphasis within a frequency spectrum form relating to a resonance of the source system. |
| <i>Fundamental Frequency</i> | Major oscillating mode of a function. |
| <i>Harmonics</i> | Strong partials at approximately integer-multiples of the fundamental frequency. |
| <i>Instrument</i> | Unified region of timbre space defining a coherent set of sounds which appear to come from the same source in some manner. |
| <i>MDS</i> | Multidimensional Scaling. Statistical procedure often used in timbre research. |
| <i>MUMS</i> | McGill University Master Samples. Archive of sounds used by a number of researchers in investigating timbre, and this work. |
| <i>Partials</i> | Time-varying frequency components. |
| <i>PCA</i> | Principal Components Analysis. Statistical procedure often used in timbre research. |
| <i>Perceptual Distance</i> | Perceived dissimilarity rated along a numerical scale. |
| <i>Peripheral Hearing System</i> | The components of the human auditory system from the outside of the head up to the nerve fibres from the cochlea to the central auditory cortex. |
| <i>SCA</i> | Sonic Character Attribute. Perceptually-based acoustical parameter of timbral variation. |
| <i>Semantic Scales</i> | Difference metrics based on verbal descriptions rather than dissimilarity. |
| <i>Spectral Aspect</i> | A general area of the spectral form as outlined in Section 2.9. |
| <i>Spectral Feature</i> | A particular component part of a the spectral form. |

| | |
|-------------------------------|--|
| <i>Spectrogram</i> | Graphical description of time-varying frequency content of a sound, where darkness corresponds to intensity, vertical axis to frequency, and horizontal axis to time. |
| <i>Stability</i> | A combination of periodicity metrics as used in Chapter 4 to control adaption of viewpoint. |
| <i>STDFT</i> | Short-Time Discrete Fourier Transform. Multiple DFTs along the length of an input sound. |
| <i>Strong Partial</i> s | Partials that have low normalised bandwidth and moderately high amplitude. |
| <i>Timbre Space</i> | A construct where distance in multidimensional space represents timbral dissimilarity between sonic entities, implicitly defining the scope of consideration in timbral experiments. |
| <i>Traditional Instrument</i> | An instrument commonly used in the Western musical tradition. |

Bibliography

- [1] ANSI, 1960. *USA Standard Acoustical Terminology (Including Mechanical Shock and Vibration)*. Technical Report, American National Standards Institute, New York. Report S1.1-1960 (R1976), Section 12.9.
- [2] B.S. Atal and S.L. Hanauer, 1971. *Speech Analysis and Synthesis by Linear Prediction of the Speech Wave*. Journal of the Acoustical Society of America, Vol.50, Pt.2, No.2, 637–655.
- [3] G.J. Balzano, 1986. *What are Musical Pitch and Timbre?* Music Perception, Vol.3, No.3, 294–314.
- [4] C.M. Barnes, 1995. *On the Psychoacoustic Design of Pitch Detection Algorithms*. DPhil Thesis, University of York.
- [5] R.J. Baron, 1987. *The Cerebral Computer: An Introduction to the Computational Structure of the Human Brain*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- [6] J.B. Barrière, Y. Potard, and P.F. Baisnée, 1985. *Models of Continuity Between Synthesis and Processing for the Elaboration and Control of Timbre Structures*. In *International Computer Music Conference (Burnaby)*, 193–198. International Computer Music Association.

-
- [7] W.T. Bartholomew, 1942. *Acoustics of Music*. Prentice Hall, Englewood Cliffs, New Jersey.
- [8] J.W. Beauchamp, March 1997. (*personal communication*). electronic mailing list.
- [9] K.G. Beauchamp, 1987. *Transforms for Engineers: A Guide to Signal Processing*. Oxford University Press, Oxford.
- [10] A.H. Benade, 1981. *Spectral Similarities of Tones from "Especially Useful" Musical Instruments (Abstract only)*. Journal of the Acoustical Society of America, Vol.69, S37. (In supplement 1).
- [11] K.W. Berger, 1964. *Some Factors in the Recognition of Timbre*. Journal of the Acoustical Society of America, Vol.36, No.10, 1888–1891.
- [12] G. von Bismarck, 1974. *Sharpness as an Attribute of the Timbre of Steady Sounds*. Acustica, Vol.30, No.3, 159–172.
- [13] G. von Bismarck, 1974. *Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes*. Acustica, Vol.30, No.3, 146–159.
- [14] P. Boulez, 1986. *Technology and the Composer*. In S. Emmerson, editor, *The Language of Electroacoustic Music*. Macmillan Press, Basingstoke.
- [15] A.S. Bregman, 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Massachusetts.
- [16] G. Bristow, editor, 1984. *Electronic Speech Synthesis; Techniques, Technology and Applications*. Granada Publishing, London.
- [17] T. Brookes, A. Tyrrell, and D. Howard, 1996. *Musical Analysis Using a Real-Time Model of Peripheral Hearing*. In *International Computer Music Conference (Hong Kong)*, 79–82. International Computer Music Association.
- [18] J.C. Brown, 1991. *Calculation of a Constant Q Spectral Transform*. Journal of the Acoustical Society of America, Vol.89, No.1, 425–434.
- [19] A. de Bruijn, 1978. *Timbre-Classification of Complex Tones*. Acustica, Vol.40, 108–114.
- [20] W. Buxton, S. Patel, W. Reeves, and R. Baecker, 1982. *Objed and the Design of Timbral Resources*. Computer Music Journal, Vol.6, No.2, 32–44.

-
- [21] G. Byford, 1991. *An Additive Synthesis System*. MA/MSc Thesis in Music Technology, University of York.
- [22] M. Campbell and C. Greated, 1987. *The Musician's Guide to Acoustics*. J.M.Dent and Sons, London.
- [23] N.R. Campbell, 1938. *Symposium : Measurement and its Importance for Philosophy*. Aristotelian Society, Vol.17. (Supplement).
- [24] S.L. Campbell, 1994. *Uni- and Multidimensional Identification of Rise Time, Spectral Slope, and Cutoff Frequency*. Journal of the Acoustical Society of America, Vol.96, No.3, 1380–1387.
- [25] E.C. Carterette, 1989. *Perception and Physiology in the Hearing of Computed Sound*. In S. Nielzén and O. Olsson, editors, *Structure and Perception of Electroacoustic Sound and Music*. Excerpta Medica, Amsterdam.
- [26] CDP, 1994. *CDP PC System Program Reference Guide by Function*. In *Composers' Desktop Project*. CDP, York.
- [27] A. Chaigne and F. Troxler, 1988. *SONATE: An Analysis/Synthesis System of Musical Sounds Based on Perceptual Data*. In *International Computer Music Conference (Cologne)*, 399–402. International Computer Music Association.
- [28] Chambers, 1991. *Chambers Science and Technology Dictionary*. Chambers, Edinburgh.
- [29] J.M. Chowning, 1977. *The Synthesis of Complex Audio Spectra by Means of Frequency Modulation*. Computer Music Journal, Vol.1, No.2, 46–54.
- [30] R. Cogan, 1984. *New Images of Musical Sound*. Harvard University Press, Cambridge, Massachusetts.
- [31] J.B. Conant, 1947. *On Understanding Science: An Historical Approach*. Yale University Press.
- [32] 1964. *Concise Oxford Dictionary of Music*. Oxford University Press, Second edition.
- [33] D. Cooper and K.C. Ng, 1996. *A Monophonic Pitch-Tracking Algorithm Based on Waveform Periodicity Determinations Using Landmark Points*. Computer Music Journal, Vol.20, No.3, 70–78.

-
- [34] P. Cosi, G. De Poli, and P. Prandoni, 1994. *Timbre Classification with Mel-Cepstrum and Neural Nets*. In *International Computer Music Conference (Aarhus)*, 42–45. International Computer Music Association.
- [35] D.P. Creasey, D.M. Howard, and A.M. Tyrrell, 1996. *The Timbral Object - An Alternative Route to the Control of Timbre Space*. In *International Computer Music Conference (Hong Kong)*, 372–374. International Computer Music Association.
- [36] R.G. Crowder, 1989. *Imagery for Musical Timbre*. *Journal of Experimental Psychology: Human Perception and Performance*, Vol.15, No.3, 472–478.
- [37] E.P. Cunningham, 1992. *Digital Filtering: An Introduction*. Houghton Mifflin Company, Boston.
- [38] I. Daubechies, 1990. *The Wavelet Transform, Time-Frequency Localisation and Signal Analysis*. *IEEE Transactions on Information Theory*, Vol.36, No.5, 961–1005.
- [39] Ph. Depalle and G. Poirot, 1991. *SVP: A Modular System for Analysis, Processing and Synthesis of Sound Signals*. In *International Computer Music Conference (Montreal)*, 161–164. International Computer Music Association.
- [40] R. Dettmer, 1995. *A Class Act: the Rise of Object-Oriented Technology*. *IEE Review*, Vol.41, No.6, 253–256.
- [41] C. Dierbach, 1983. *Some Initial Ideas on the Control of Digital Sound Synthesis Through AI Techniques*. In *International Computer Music Conference (New York)*, 235–251. International Computer Music Association.
- [42] C. Dodge and T.A. Jerse, 1985. *Computer Music: Synthesis, Composition and Performance*. Schirmer Books, London.
- [43] S. Donnadiou, S. McAdams, and S. Winsberg, 1994. *Caractérisation du Timbre des Son Complexes. I. Analyse Multidimensionnelle*. *Journal de Physique IV*, Vol.4, 593–596. (Third French Conference on Acoustics (Toulouse), Colloque C5, translated by E.S.Creasey).
- [44] S. Donnadiou, S. McAdams, and S. Winsberg, 1996? *Effects of Context Change on Timbre Perception*. from http://www.cnmat.berkeley.edu/Abshtml/Donnadiou_et_al.html.

-
- [45] S. Dubnov and N. Tishby, 1995. *Clustering of Musical Sounds using Polyspectral Distance Measures*. In *International Computer Music Conference (Banff)*, 460–466. International Computer Music Association.
- [46] S. Dubnov, N. Tishby, and D. Cohen, 1995. *Hearing Beyond the Spectrum*. *Journal of New Music Research*, Vol.24, 342–368.
- [47] D.A. Eddins and D.M. Green, 1995. *Temporal Integration and Temporal Resolution*. In B.C.J. Moore, editor, *Hearing*. Academic Press, London.
- [48] M.D. Edgington and J.A.S. Angus, 1992. *A Transform Method for Generating Perceptually Biased Spectrograms*. *Proceedings of the Institute of Acoustics*, Vol.14, Pt.6, 569–576.
- [49] D.P.W. Ellis and B.L. Vercoe, 1991. *A Wavelet-Based Sinusoid Model of Sound for Auditory Signal Separation*. In *International Computer Music Conference (Montreal)*, 86–89. International Computer Music Association.
- [50] S. Emmerson, 1986. *The Relation of Language to Materials*. In S. Emmerson, editor, *The Language of Electroacoustic Music*. Macmillan Press, Basingstoke.
- [51] R. Erickson, 1975. *Sound Structure in Music*. University of California Press, Berkeley.
- [52] R. Ethington and B. Punch, 1994. *Sea Wave: A System for Musical Timbre Description*. *Computer Music Journal*, Vol.18, No.1, 30–39.
- [53] E.F. Evans, 1982. *Basic Physics and Psychophysics of Sound*. In H.B. Barlow and J.D. Mollon, editors, *The Senses*, 239–250. Cambridge University Press.
- [54] E.F. Evans, 1982. *Functions of the Auditory System*. In H.B. Barlow and J.D. Mollon, editors, *The Senses*, 307–332. Cambridge University Press.
- [55] F. Fallside and W.A. Woods, editors, 1985. *Computer Speech Processing*. Prentice-Hall International, Englewood Cliffs, New Jersey.
- [56] B. Feiten, R. Frank, and T. Ungvary, 1991. *Organisation of Sounds with Neural Nets*. In *International Computer Music Conference (Montreal)*, 441–444. International Computer Music Association.

-
- [57] K. Fitz and L. Haken, 1996. *Sinusoidal Modeling and Manipulation Using Lemur*. Computer Music Journal, Vol.20, No.4, 44–59.
- [58] K. Fitz, W. Walker, and L. Haken, 1992. *Extending the McAulay-Quatieri Analysis for Synthesis with a Limited Number of Oscillators*. In *International Computer Music Conference (San Jose)*, 381–382. International Computer Music Association.
- [59] H. Fletcher, 1934. *Loudness, Pitch and the Timbre of Musical Tones and their Relation to the Intensity, the Frequency and the Overtone Structure*. Journal of the Acoustical Society of America, Vol.6, 59–69.
- [60] N.H. Fletcher and T.D. Rossing, 1991. *The Physics of Musical Instruments*. Springer-Verlag, New York.
- [61] P. Forrest, 1992. *The Moog Series III : The Start of Something Big?* Music Technology, Vol.6, No.10, 62–65.
- [62] D.L. Freed and W.L. Martens, 1986. *Deriving Psychophysical Relations for Timbre*. In *International Computer Music Conference (The Hague)*, 393–405. International Computer Music Association.
- [63] M.D. Freedman, 1967. *Analysis of Musical Instrument Tones*. Journal of the Acoustical Society of America, Vol.41, No.4, 793–806.
- [64] S. De Furia and J. Scacciaferro, 1987. *The Sampling Book*. Third Earth Publishing, Pompton Lakes, New Jersey.
- [65] D. Gabor, 1947. *Acoustical Quanta and the Theory of Hearing*. Nature, Vol.159, No.4044, 591–594.
- [66] W.W. Gaver, 1993. *How Do We Hear in the World?: Explorations in Ecological Acoustics*. Ecological Psychology, Vol.5, No.4, 285–313.
- [67] W.W. Gaver, 1993. *What in the World Do We Hear?: An Ecological Approach to Auditory Event Perception*. Ecological Psychology, Vol.5, No.1, 1–29.
- [68] W.H. George, 1954. *A Sound Reversal Technique Applied to the Study of Tone Quality*. Acustica, Vol.4, 224–225.

-
- [69] B. Gold and L. Rabiner, 1969. *Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain*. Journal of the Acoustical Society of America, Vol.46, Pt.2, No.2, 442–448.
- [70] J.W. Gordon and J.M. Grey, 1978. *Perception of Spectral Modifications on Orchestral Instrument Tones*. Computer Music Journal, Vol.2, No.1, 24–31.
- [71] J.M. Grey, 1975. *An Exploration of Musical Timbre Using Computer-Based Techniques For Analysis, Synthesis and Perceptual Scaling*. PhD Thesis, Stanford University, California.
- [72] J.M. Grey, 1977. *Multidimensional Perceptual Scaling of Musical Timbres*. Journal of the Acoustical Society of America, Vol.61, No.5, 1270–1277.
- [73] J.M. Grey, 1978. *Timbre Discrimination in Musical Patterns*. Journal of the Acoustical Society of America, Vol.64, No.2, 467–472.
- [74] J.M. Grey and J.W. Gordon, 1978. *Perceptual Effects of Spectral Modifications on Musical Tones*. Journal of the Acoustical Society of America, Vol.63, No.5, 1493–1500.
- [75] J. Hajda, March 1997. (*personal communication*). electronic mailing list.
- [76] J.M. Hajda, 1996? *Issues in Timbre Research*. from <http://www.cnmat.berkeley.edu/Abshtml/Hajda.html>.
- [77] D.E. Hall, 1991. *Musical Acoustics*. Brooks/Cole Publishing, Pacific Grove, California, *Second* edition.
- [78] J.W. Hall, J.H. Grose, and L. Mendoza, 1995. *Across-Channel Processes in Masking*. In B.C.J.Moore, editor, *Hearing*. Academic Press, London.
- [79] L.C. Hamilton, 1990. *Modern Data Analysis: A First Course in Applied Statistics*. Brooks/Cole Publishing, Pacific Grove, California.
- [80] S. Handel, 1995. *Timbre Perception and Auditory Object Identification*. In B.C.J.Moore, editor, *Hearing*. Academic Press, London.
- [81] F.J. Harris, 1978. *On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform*. Proceedings of the IEEE, Vol.66, No.1, 51–83.

-
- [82] J. Harvey, 1986. *The Mirror of Ambiguity*. In S. Emmerson, editor, *The Language of Electroacoustic Music*. Macmillan Press, Basingstoke.
- [83] H.L.F. von Helmholtz, 1877. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Dover Publications, New York, *Second edition*. Translated by A.J.Ellis (1954).
- [84] W. Hess, 1983. *Pitch Determination of Speech Signals; Algorithms and Devices*. Springer-Verlag, Berlin.
- [85] R.D. Hill, 1991. *The Cro-Magnon Advanced Additive Analysis/Synthesis System*. In *International Computer Music Conference (Montreal)*, 169–176. International Computer Music Association.
- [86] J.N. Holmes, 1988. *Speech Synthesis and Recognition*. Von Nostrand Reinhold, Wokingham.
- [87] C. Hourdin, G. Charbonneau, and T. Moussa, 1997. *A Multidimensional Scaling Analysis of Musical Instruments' Time-Varying Spectra*. *Computer Music Journal*, Vol.21, No.2, 40–55.
- [88] D.M. Howard and J. Angus, 1996. *Acoustics and Psychoacoustics*. Focal Press, Oxford.
- [89] J.H. Howard and J.A. Ballas, 1981. *Feature Selection in Auditory Perception*. In D.J. Getty and J.H. Howard, editors, *Auditory and Visual Pattern Recognition*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- [90] E.C. Ifeachor and B.W. Jervis, 1993. *Digital Signal Processing; A Practical Approach*. Addison Wesley, Wokingham, England.
- [91] IRCAM, 1995. *La Psychoacoustique*. (Translated by E.S.Creasey), from <http://www.ircam.fr/equipes/psychoacoustique.html>.
- [92] P. Iverson and C.L. Krumhansl, 1989. *Pitch and Timbre Interaction in Isolated Tones and in Sequences (Abstract only)*. *Journal of the Acoustical Society of America*, Vol.86, S58. (In supplement 1).
- [93] P. Iverson and C.L. Krumhansl, 1993. *Isolating the Dynamic Attributes of Musical Timbre*. *Journal of the Acoustical Society of America*, Vol.94, No.5, 2595–2603.

-
- [94] D.A. Jaffe and J.O. Smith, 1983. *Extensions of the Karplus-Strong Plucked-String Algorithm*. Computer Music Journal, Vol.7, No.2, 56–69.
- [95] C. Jansen, 1992. *Sine Circuitu; Real-Time Analysis, Manipulation and (Re)Synthesis*. In *International Computer Music Conference (San Jose)*, 451–452. International Computer Music Association.
- [96] J. Jeans, 1961. *Science and Music*. Cambridge University Press.
- [97] D. Keane, 1986. *At the Threshold of an Aesthetic*. In S. Emmerson, editor, *The Language of Electroacoustic Music*. Macmillan Press, Basingstoke.
- [98] D. Keislar, T. Blum, J. Wheaton, and E. Wold, 1995. *Audio Analysis for Content-Based Retrieval*. In *International Computer Music Conference (Banff)*, 199–202. International Computer Music Association.
- [99] O. Kitamura, S. Namba, and R. Matsumoto, 1968. *Factor Analytical Research of Tone Color*. In *Reports of the 6th International Congress on Acoustics (Tokyo)*, 117–120.
- [100] J. Krimphoff, S. McAdams, and S. Winsberg, 1994. *Caractérisation du Timbre des Son Complexes. II. Analyses Acoustique et Quantification Psychophysique*. Journal de Physique IV, Vol.4, 625–628. (Third French Conference on Acoustics (Toulouse), Colloque C5, translated by E.S.Creasey).
- [101] R. Kronland-Martinet, 1988. *The Wavelet Transform for Analysis, Synthesis, and Processing of Speech and Music Sounds*. Computer Music Journal, Vol.12, No.4, 11–20.
- [102] C.L. Krumhansl, 1989. *Why is Musical Timbre So Hard to Understand?* In S. Nielzén and O. Olsson, editors, *Structure and Perception of Electroacoustic Sound and Music*. Excerpta Medica, Amsterdam.
- [103] W.B. Kuhn, 1990. *A Real-Time Pitch Recognition Algorithm for Music Applications*. Computer Music Journal, Vol.14, No.3, 60–71.
- [104] C. Kussmaul, 1991. *Applications of the Wavelet Transform at the Level of the Pitch Contour*. In *International Computer Music Conference (Montreal)*, 483–486. International Computer Music Association.

-
- [105] S. Lakatos and S. McAdams, 1996? *Scaling of Harmonic and Percussive Timbres: Perceptual Spaces and Verbal Attributes*. from <http://www.cnmat.berkeley.edu/Abshtml/McAdams&Lakatos.html>.
- [106] C.J. Langmead, 1995. *PAST Manual; Version 1.6*. Dartmouth College, Department of Electro-Acoustic Music, Hanover, New Hampshire.
- [107] C.J. Langmead, 1995. *A Theoretical Model of Timbre Perception Based on Morphological Representations of Time-Varying Spectra*. M.A. Thesis, Dartmouth College, Department of Electro-Acoustic Music, Hanover, New Hampshire.
- [108] K. Lassfolk, 1996. *Simulation of Electron Tube Audio Circuits*. In *International Computer Music Conference (Hong Kong)*, 222–223. International Computer Music Association.
- [109] R.G. Laughlin, B.D. Truax, and B.V. Funt, 1990. *Synthesis of Acoustic Timbres Using Principle Component Analysis*. In *International Computer Music Conference (Glasgow)*, 95–99. International Computer Music Association.
- [110] M.R. Leek, 1987. *Directed Attention in Complex Sound Perception*. In W.A. Yost and C.S. Watson, editors, *Auditory Processing of Complex Sounds*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- [111] F. Lerdahl, 1987. *Timbral Hierarchies*. *Contemporary Music Review*, Vol.2, 135–160.
- [112] W.H. Lichte, 1941. *Attributes of Complex Tones*. *Journal of Experimental Psychology*, Vol.28, No.6, 455–480.
- [113] J.C.R. Licklider, 1951. *Basic Correlates of the Auditory Stimulus*. In S.S. Stevens, editor, *Handbook of Experimental Psychology*. John Wiley and Sons, New York.
- [114] J.S. Lim and A.V. Oppenheim, 1988. *Advanced Topics in Signal Processing*. Prentice Hall, Englewood Cliffs, New Jersey.
- [115] Y-O. Lo, 1986. *Techniques of Timbral Interpolation*. In *International Computer Music Conference (The Hague)*, 241–247. International Computer Music Association.
- [116] P. Lundén, 1993. *Knowledge Representation of Sounds and Sonic-Structures Based on Constraints and Multiple Inheritance*. In *International Computer Music Conference (Tokyo)*, 369–371. International Computer Music Association.

-
- [117] N. Magnus, 1994. *Creating Analogue Sounds on Digital Synths: Part 3*. Sound On Sound, Vol.9, No.5, 46–48.
- [118] T.F. Malet, 1993. *An Investigation into a New Representation of Music Using Harmonic Analysis*. PhD Thesis, University of Sheffield.
- [119] B.F.J. Manly, 1994. *Multivariate Statistical Methods: A Primer*. Chapman & Hall, London, *Second* edition.
- [120] W.L. Martens, 1985. *Palette: An Environment for Developing an Individualized Set of Psychophysically Scaled Timbres*. In *International Computer Music Conference (Burnaby)*, 355–365. International Computer Music Association.
- [121] D.W. Massaro, 1987. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- [122] S. McAdams and A. Bregman, 1979. *Hearing Musical Streams*. Computer Music Journal, Vol.3, No.4, 26–43,60,63.
- [123] S. McAdams and J-C. Cunible, 1992. *Perception of Timbral Analogies*. Philosophical Transactions of the Royal Society of London, Vol.336, Pt.B, 383–389.
- [124] R.J. McAulay and T.F. Quatieri, 1986. *Speech Analysis/Synthesis Based on a Sinusoidal Representation*. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-34, No.4, 744–754.
- [125] W.F. McGee and P. Merkley, 1989. *Real-Time Acoustic Analysis of Polyphonic Music*. In *International Computer Music Conference (Ohio)*, 199–202. International Computer Music Association.
- [126] T. McLaughlin, 1992. *On the Attack*. Music Technology, Vol.6, No.2, 64.
- [127] K. McMillen, D.L. Wessel, and M. Wright, 1994. *The ZIPI Music Parameter Description Language*. Computer Music Journal, Vol.18, No.4, 52–73.
- [128] M. McNabb, 1986. *Computer Music: Some Aesthetic Considerations*. In S. Emmerson, editor, *The Language of Electroacoustic Music*. Macmillan Press, Basingstoke.
- [129] R.D. Melara and L.E. Marks, 1990. *Interaction Among Auditory Dimensions: Timbre, Pitch and Loudness*. Perception and Psychophysics, Vol.48, No.2, 169–178.

-
- [130] A. Melka, 1994. *Methodological Approaches to the Investigation of Musical Timbre*. Journal de Physique IV, Vol.4, 569–576. (Third French Conference on Acoustics (Toulouse), Colloque C5).
- [131] J.R. Miller and E.C. Carterette, 1975. *Perceptual Space for Musical Structures*. Journal of the Acoustical Society of America, Vol.58, No.3, 711–720.
- [132] E.R. Miranda, 1994. *An Artificial Intelligence Approach to Sound Design*. from http://turandot.music.ed.ac.uk/pgregs/eduardo/information_miranda.html.
- [133] E.R. Miranda, 1994. *From Symbols to Sound : AI-based Investigation of Sound Synthesis*. Contemporary Music Review, Vol.10, Pt.2, 211–232.
- [134] B.C.J. Moore, 1995. *Frequency Analysis and Masking*. In B.C.J. Moore, editor, *Hearing*. Academic Press, London.
- [135] B.C.J. Moore and B.R. Glasberg, 1983. *Suggested Formulae for Calculating Auditory-Filter Bandwidths and Excitation Patterns*. Journal of the Acoustical Society of America, Vol.74, No.3, 750–753.
- [136] J.A. Moorer, J. Grey, and J. Snell, 1977. *Lexicon of Analyzed Tones, Part 1: A Violin Tone*. Computer Music Journal, Vol.1, No.2, 39–45.
- [137] J.A. Moorer, J. Grey, and J. Strawn, 1977. *Lexicon of Analyzed Tones, Part 2: Clarinet and Oboe Tones*. Computer Music Journal, Vol.1, No.3, 12–29.
- [138] J.A. Moorer, J. Grey, and J. Strawn, 1978. *Lexicon of Analyzed Tones, Part 3: The Trumpet*. Computer Music Journal, Vol.2, No.2, 23–31.
- [139] K. Nordenstreng, 1969. *The Perception of Complex Sounds: Semantic Differential Attributes of Speech and Music*. In J. Järvinen, editor, *Contemporary Research in Psychology of Perception*. Werner Söderström Osakeyhtiö, Porvoo, Helsinki.
- [140] M.J. Norušis, 1994. *SPSS 6.1 Base System User's Guide Part 2*. SPSS Inc.
- [141] M.J. Norušis, 1994. *SPSS Professional Statistics 6.1*. SPSS Inc.
- [142] V.N. Nosulenko, E.S. Samoylenko, and S. McAdams, 1994. *L'analyse de Descriptions Verbales dans L'étude des Comparaisons de Timbre Musicaux*. Journal de Physique IV, Vol.4, 637–640. (Third French Conference on Acoustics (Toulouse), Colloque C5, translated by E.S. Creasey).

-
- [143] C.H. Null and F.W. Young, 1981. *Auditory Perception: Recommendations for a Computer Assisted Experimental Paradigm*. In D.J. Getty and J.H. Howard, editors, *Auditory and Visual Pattern Recognition*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.
- [144] A.V. Oppenheim, 1969. *Speech Analysis/Synthesis System Based on Homomorphic Filtering*. *Journal of the Acoustical Society of America*, Vol.45, No.2, 458–465.
- [145] A.V. Oppenheim and R.W. Schafer, 1975. *Digital Signal Processing*. Prentice Hall, Englewood Cliffs, New Jersey.
- [146] D.V. Oppenheim, T. Anderson, and R. Kirk, 1993. *Perceptual Parameters - Their Specification, Scoring and Control within Two Software Composition Systems*. In *International Computer Music Conference (Tokyo)*, 421–424. International Computer Music Association.
- [147] N. Osaka, 1995. *Timbre Interpolation of Sounds Using a Sinusoidal Model*. In *International Computer Music Conference (Banff)*, 408–411. International Computer Music Association.
- [148] 1989. *The Oxford English Dictionary, Volume XVIII*. Clarendon Press (Oxford University Press), *Second* edition.
- [149] C. Padgham, 1986. *The Scaling of the Timbre of the Pipe Organ*. *Acustica*, Vol.60, No.3, 189–204.
- [150] A.R. Palmer, 1995. *Neural Signal Processing*. In B.C.J. Moore, editor, *Hearing*. Academic Press, London.
- [151] R.D. Patterson, 1989. *Timbre and Tone Height (Abstract only)*. *Journal of the Acoustical Society of America*, Vol.86, S58. (In supplement 1).
- [152] P-G. Pérès-Labourdette, 1990. *A Study of FM Synthesis and Neural Networks*. MA/MSc Thesis in Music Technology, University of York.
- [153] J.R. Pierce, 1983. *The Science of Musical Sound*. Scientific American Books, New York.
- [154] M.A. Pitt, 1995. *Evidence for a Central Representation of Instrument Timbre*. *Perception and Psychophysics*, Vol.57, No.1, 43–55.

-
- [155] M.A. Pitt and R.G. Crowder, 1992. *The Role of Spectral and Dynamic Cues in Imagery for Musical Timbre*. *Journal of Experimental Psychology: Human Perception and Performance*, Vol.18, No.3, 728–738.
- [156] R. Plomp, 1971. *Timbre as a Multidimensional Attribute of Complex Tones*. In R. Plomp and G.F. Smoorenburg, editors, *Frequency Analysis and Periodicity Detection in Hearing*, 397–414. Sijthoff, Leiden.
- [157] R. Plomp, 1976. *Aspects of Tone Sensation: A Psychophysical Study*. Academic Press, London.
- [158] R. Plomp and W.J.M. Levelt, 1965. *Tonal Consonance and Critical Bandwidth*. *Journal of the Acoustical Society of America*, Vol.38, 548–560.
- [159] R. Plomp and H.J.M. Steeneken, 1969. *Effect of Phase on the Timbre of Complex Tones*. *Journal of the Acoustical Society of America*, Vol.46, Pt.2, No.2, 409–421.
- [160] G. De Poli and P. Tonella, 1993. *Self-Organising Neural Network and Grey's Timbre Space*. In *International Computer Music Conference (Tokyo)*, 260–263. International Computer Music Association.
- [161] H.F. Pollard, 1988. *Feature Analysis of Musical Sounds*. *Acustica*, Vol.65, 232–244.
- [162] H.F. Pollard, 1990. *Timbre Assessment*. *Acoustics Australia*, Vol.18, No.1, 19–23.
- [163] H.F. Pollard, 1990. *Timbre Measurement*. *Acoustics Australia*, Vol.18, No.3, 65–69.
- [164] H.F. Pollard and E.V. Jansson, 1982. *Analysis and Assessment of Musical Starting Transients*. *Acustica*, Vol.51, 249–262.
- [165] H.F. Pollard and E.V. Jansson, 1982. *A Tristimulus Method for the Specification of Musical Timbre*. *Acustica*, Vol.51, 162–171.
- [166] Y. Potard, P.F. Baisnée, and J.B. Barrière, 1986. *Experimenting With Models of Resonance Produced by a New Technique for the Analysis of Impulsive Sounds*. In *International Computer Music Conference (The Hague)*, 269–274. International Computer Music Association.
- [167] R.L. Pratt and P.E. Doak, 1976. *A Subjective Rating Scale for Timbre*. *Journal of Sound and Vibration*, Vol.43, No.3, 317–328.

-
- [168] A. Preis, 1984. *An Attempt to Describe the Parameter Determining the Timbre of Steady-State Harmonic Complex Tones*. *Acustica*, Vol.55, No.1, 1–13.
- [169] L.R. Rabiner and B. Gold, 1975. *Theory and Application of Digital Signal Processing*. Prentice Hall, Englewood Cliffs, New Jersey.
- [170] L.R. Rabiner and R.W. Schafer, 1978. *Digital Processing of Speech Signals*. Prentice Hall International, Englewood Cliffs, New Jersey.
- [171] R.A. Rasch and R. Plomp, 1982. *The Perception of Musical Tones*. In D. Deutsch, editor, *The Psychology of Music*. Academic Press, London.
- [172] J-C. Risset and M.V. Mathews, 1969. *Analysis of Musical-Instrument Tones*. *Physics Today*, Vol.22, No.2, 23–30.
- [173] J-C. Risset and D.L. Wessel, 1982. *Exploration of Timbre by Analysis and Synthesis*. In D. Deutsch, editor, *The Psychology of Music*. Academic Press, London.
- [174] T.D. Rossing, 1990. *The Science of Sound*. Addison Wesley Publishing, Reading, Massachusetts, *Second* edition.
- [175] E.L. Saldanha and J.F. Corso, 1964. *Timbre Cues and the Identification of Musical Instruments*. *Journal of the Acoustical Society of America*, Vol.36, No.11, 2021–2026.
- [176] G.J. Sandell, 1991. *A Library of Orchestral Instrument Spectra*. In *International Computer Music Conference (Montreal)*, 98–101. International Computer Music Association.
- [177] G.J. Sandell and W.L. Martens, 1992. *Prototyping and Interpolation of Multiple Musical Timbres Using Principle Component-Based Synthesis*. In *International Computer Music Conference (San Jose)*, 34–37. International Computer Music Association.
- [178] G.J. Sandell and W.L. Martens, 1995. *Perceptual Evaluation of Principle-Component-Based Synthesis of Musical Timbres*. *Journal of the Audio Engineering Society*, Vol.43, No.12, 1013–1028.
- [179] J.F. Schouten, 1968. *The Perception of Timbre*. In *Reports of the 6th International Congress on Acoustics (Tokyo)*, 89–90.

-
- [180] C.E. Seashore, 1938. *Psychology of Music*. Dover Publications, New York. 1967 Reprint.
- [181] X. Serra and J. Smith, 1990. *Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition*. *Computer Music Journal*, Vol.14, No.4, 12–24.
- [182] X. Serra and J.O. Smith, 1989. *Spectral Modeling Synthesis*. In *International Computer Music Conference (Ohio)*, 281–283. International Computer Music Association.
- [183] Z. Settel and C. Lippe, 1994. *Real Time Musical Applications using FFT-Based Resynthesis*. In *International Computer Music Conference (Aarhus)*, 338–343. International Computer Music Association.
- [184] S.A. Shamma, S. Vranić, and P. Wiser, 1992. *Spectral Gradient Columns in Primary Auditory Cortex: Physiological and Psychological Correlates*. In Y. Cazals, L. Demany, and K. Horner, editors, *Auditory Physiology and Perception: Advances in the Biosciences, Volume 83*. Pergamon Press, Oxford.
- [185] P.G. Singh and I.J. Hirsh, 1992. *Influence of Spectral Locus and F0 Changes on the Pitch and Timbre of Complex Tones*. *Journal of the Acoustical Society of America*, Vol.92, No.5, 2650–2661.
- [186] L. Sjöberg and C. Thorslund, 1979. *A Classificatory Theory of Similarity*. *Psychological Research*, Vol.40, 223–247.
- [187] M. Slaney, 1993. *An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank*. Technical Report 35, Apple Computer, Cupertino, California.
- [188] A.W. Slawson, 1968. *Vowel Quality and Musical Timbre as Functions of Spectrum Envelope and Fundamental Frequency*. *Journal of the Acoustical Society of America*, Vol.43, No.1, 87–101.
- [189] W. Slawson, 1984. *Operations on Timbre: Perspectives and Problems*. In *International Computer Music Conference (Paris)*, 167–171. International Computer Music Association.
- [190] W. Slawson, 1985. *Sound Color*. University of California Press, London.

-
- [191] W. Slawson, 1986. *Sound-Color Dynamics*. Perspectives of New Music, Vol.25, 156–181.
- [192] W. Slawson, 1989. *Sound Structure and Musical Structure: The Role of Sound Color*. In S. Nielzén and O. Olsson, editors, *Structure and Perception of Electroacoustic Sound and Music*. Excerpta Medica, Amsterdam.
- [193] D. Smalley, 1986. *Spectro-Morphology and Structuring Processes*. In S. Emmerson, editor, *The Language of Electroacoustic Music*. Macmillan Press, Basingstoke.
- [194] D. Smalley, 1994. *Defining Timbre - Refining Timbre*. Contemporary Music Review, Vol.10, Pt.2, 35–48.
- [195] J.O. Smith and X. Serra, 1987. *PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation*. In *International Computer Music Conference (San Francisco)*, 290–297. International Computer Music Association.
- [196] L.N. Solomon, 1958. *Semantic Approach to the Perception of Complex Sounds*. Journal of the Acoustical Society of America, Vol.30, No.5, 421–425.
- [197] L.N. Solomon, 1959. *Search for Physical Correlates to Psychological Dimensions of Sounds*. Journal of the Acoustical Society of America, Vol.31, No.4, 492–497.
- [198] S.S. Stevens, 1975. *Psychophysics : Introduction to its Perceptual, Neural, and Social Prospects*. John Wiley and Sons, London.
- [199] E. Tellman, L. Haken, and B. Holloway, 1995. *Timbre Morphing of Sounds With Unequal Numbers of Features*. Journal of the Audio Engineering Society, Vol.43, No.9, 678–689.
- [200] P. Toiviainen, 1996. *Optimizing Auditory Images and Distance Metrics for Self-Organizing Timbre Maps*. Journal of New Music Research, Vol.25, 1–30.
- [201] P. Toiviainen, M. Kaipainen, and J. Louhivuori, 1995. *Musical Timbre: Similarity Ratings Correlate with Computational Feature Space Distances*. Journal of New Music Research, Vol.24, 282–298.
- [202] J.T. Tou, 1981. *A Feature-Extraction Approach to Auditory Pattern Recognition*. In D.J. Getty and J.H. Howard, editors, *Auditory and Visual Pattern Recognition*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

-
- [203] S. Trask, 1990. *S770*. Music Technology, Vol.4, No.8, 56–58.
- [204] S Trask, 1992. *JV80 Synth*. Music Technology, Vol.6, No.6, 60–66.
- [205] A. Tversky, 1977. *Features of Similarity*. Psychological Review, Vol.84, No.4, 327–352.
- [206] H. Uematsu, K. Ozawa, Y. Susuki, and T. Sone, 1996. *A Consideration on the Difference Limen for Timbre of Complex Tones Consisting of Higher Harmonics*. Journal of the Acoustical Society of Japan (E), Vol.17, No.2, 105–108.
- [207] P.P. Vaidyanathan, 1987. *Quadrature Mirror Filter Banks, M-Band Extensions and Perfect-Reconstruction Techniques*. IEEE ASSP Magazine, Vol.4, No.3, 4–19.
- [208] P.P. Vaidyanathan, 1990. *Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial*. Proceedings of the IEEE, Vol.78, No.1, 56–93.
- [209] J. Vandenheede and J. Harvey, 1985. *Identity and Ambiguity: The Construction and Use of Timbral Transitions and Hybrids*. In *International Computer Music Conference (Burnaby)*, 97–102. International Computer Music Association.
- [210] N.J. Versfeld and A.J.M. Houtsma, 1992. *Spectral Shape Discrimination of Two-tone Complexes*. In Y. Cazals, L. Demany, and K. Horner, editors, *Auditory Physiology and Perception: Advances in the Biosciences, Volume 83*. Pergamon Press, Oxford.
- [211] R. Vertegaal and E. Bonis, 1994. *ISEE: An Intuitive Sound Editing Environment*. Computer Music Journal, Vol.18, No.2, 21–29.
- [212] R. Vertegaal and B. Eaglestone, 1996. *Comparison of Input Devices in an ISEE Direct Timbre Manipulation Task*. Interacting With Computers, Vol.8, No.1, 13–30.
- [213] S. Waters, 1994. *Timbre Composition: Ideology, Metaphor and Social Process*. Contemporary Music Review, Vol.10, Pt.2, 129–134.
- [214] J.C. Webster, A. Carpenter, and M.M. Woodhead, 1968. *Identifying Meaningless Tonal Complexes*. Journal of the Acoustical Society of America, Vol.44, No.2, 606–609.
- [215] J.C. Webster, M.M. Woodhead, and A. Carpenter, 1970. *Perceptual Constancy in Complex Sound Identification*. British Journal of Psychology, Vol.61, No.4, 481–489.

-
- [216] L. Wedin, 1972. *A Multidimensional Study of Perceptual-Emotional Qualities in Music*. Scandinavian Journal of Psychology, Vol.13, 241–257.
- [217] L. Wedin and G. Goude, 1972. *Dimension Analysis of the Perception of Instrumental Timbre*. Scandinavian Journal of Psychology, Vol.13, 228–240.
- [218] D.L. Wessel, 1979. *Timbre Space as a Musical Control Structure*. Computer Music Journal, Vol.3, No.2, 45–52.
- [219] P. White, 1992. *The Gentle Art of EQ*. Recording Musician, Vol.1, No.3, 50–55.
- [220] P. White, 1994. *The Boom is Back*. Sound On Sound, Vol.9, No.5, 52–53.
- [221] P. White, 1994. *Sound Foundation: Part 1*. Sound On Sound, Vol.9, No.4, 60–67.
- [222] P. White, 1994. *Sound Foundation: Part 2*. Sound On Sound, Vol.9, No.5, 70–78.
- [223] J.P. Wilson, 1992. *Cochlear Mechanics*. In Y. Cazals, L. Demany, and K. Horner, editors, *Auditory Physiology and Perception: Advances in the Biosciences, Volume 83*. Pergamon Press, Oxford.
- [224] T. Wishart, 1986. *Sound Symbols and Landscapes*. In S. Emmerson, editor, *The Language of Electroacoustic Music*. Macmillan Press, Basingstoke.
- [225] T. Wishart, 1988. *The Composition of Vox-5*. Computer Music Journal, Vol.12, No.4, 21–27.
- [226] T. Wishart, 1994. *Audible Design; A Plain and Easy Introduction to Practical Sound Composition*. Orpheus The Pantomime, York.
- [227] I.H. Witten, 1982. *Principles of Computer Speech*. Academic Press, London.
- [228] T. Wright, 1993. *An Investigation into the Use of Neural Networks to Map an FM Synthesis Parameter onto Subjective Timbral Descriptions*. MA/MSc Thesis in Music Technology, University of York.
- [229] G.K. Yates, 1995. *Cochlear Structure and Function*. In B.C.J. Moore, editor, *Hearing*. Academic Press, London.

Index

- abstract, 3
- acknowledgements, 13
- acoustical couplings, 30, 177
- adaptor stimulus, 93
- amplitude envelope, 67, 73
- amplitude envelope modes, 191
- amplitude gradient, 191
- analysis-synthesis
 - combining bin data based on
 - stability, 167
 - conclusions, 169
 - extracting partial tracks, 167
 - method and results, 154
 - multiple-resolution analysis, 164
 - previous spectral systems, 149
 - spectral systems background, 147
 - stability analysis, 158
 - stability concepts, 157
 - stability/amplitude analysis in bins,
 - 165
 - synthesis scheme, 168
 - technique employed, 152
 - testing procedure, 155
- analysis-synthesis model, 146
- applications, 18, 44
- attack, 66, 72, 190
- auditory path, 16, 147
- AXESDIST, 200, 206, 258
 - metrics, 201
 - testing procedure, 201
- breakpoints, 191
- comb technique, 182
- comparison with similar studies, 224
- composite model form, 83
- conclusions, overall, 228
- confusion, source, 126, 202
- contributions to knowledge, 20, 234
- correlation, 114, 174, 203, 253
- decay, 67, 73
- declaration, 15
- dimensionality, 19, 44, 46, 99, 173, 199,
 - 200, 203, 205, 206, 216, 217, 227
 - conclusions, 223
 - definition, 200, 220
 - factors affecting, 48, 217

-
- Discrete Fourier Transform (DFT), 147, 150, 153, 164
 - duration, 29, 30
 - elements of sound, 17, 29
 - equalising sounds, 30
 - equipment and software, 14
 - extracted feature set, 195
 - feature analysis
 - 3D hierarchical decomposition, 206
 - 3D high level distinctions, 207
 - 3D instrument group distinctions, 211
 - 3D lower level distinctions, 215
 - comparison with similar studies, 224
 - correlation with scaled perceptual dimensions, 203
 - details, 202
 - dimensionality and noise effects, 217
 - dimensionality conclusions, 223
 - dimensionality high level example, 219
 - dimensionality limited feature set examples, 221
 - dimensionality limited stimulus set example, 220
 - discussion, 224
 - objectives, 202
 - overview, 196
 - similarities with perceptual results, 208
 - stimulus groups, 206
 - technique, 200
 - time-varying characteristic codes, 198
 - time-varying form characterisation, 197
 - feature extraction
 - amplitude envelope, 189
 - amplitude envelope form static measures algorithms, 190
 - areas of consideration, 175
 - details, 177
 - fundamental frequency algorithm, 178
 - logical relationships with perception, 226
 - major elements, 173
 - spectral shape time-varying measures algorithms, 192
 - static and time-varying, 174
 - strong partial characteristics time-varying measures algorithms, 187
 - summary, 195
 - testing procedure, 193
 - feature extraction/analysis, 170
 - background, 171
 - concepts to be established, 173
 - conclusions, 225
 - financial support, 14
 - frequency patterning, 68, 73
 - fundamental frequency, 178
 - further work, 236
 - glossary, 261
 - ambiguous terms, 21
 - harmonic form, 68, 73, 178, 182, 187, 188
-

-
- hierarchical decomposition, 54, 173, 175,
206, 216, 226
- high level timbral concepts, 86
- hypothesis, 2
confirmation, 238
details, 19
- inharmonic form, 68, 73, 188
- input stimuli, 244
- instrument, 42
definition, 39
traditional, 40
- instrument space, 32, 107
- intersection of timbral characteristics, 144
- introduction, 16
- levels of detail, 26, 28, 44
- loudness, 25, 29, 30
- manipulable form, 151, 242
- mathematical methods, 251
basic methods, 252
group discrimination, 257
multivariate scaling, 254
- McGill University Master Samples
(MUMS), 43, 244
- model structure, 54, 99
- modelling problem, 17
- motivation and aims, 17
- multidimensional scaling (MDS), 62, 64,
200, 256
- neural networks, 64
- noise, 69, 74
- novel aspects, 87, 143, 169, 225, 234
- objectivity, 45, 174, 175, 202, 209, 216,
226
- onset, 66, 72, 190
- partial tracks, 151, 167, 168
- past research
limitations, 81
reliability, 84
- peaks and troughs, 191
- perception
classificatory, 95, 139
context and directed attention, 93
continuous, 95
judging similarity, 96
listening types, 93
prior knowledge, 99
relationship to acoustical forms, 58,
171, 197
source identification, 98
structure of information, 91
- perception mechanisms, 91
- perceptual study, 88
concepts to be established, 89
conclusions, 143
correlation coefficients, 114
limitations, 141
major characteristics, 101
modal responses, 121
questionnaire format, 102
reasons for, 89
response distributions, 113
results analysis, 108
results format information, 108
scaled stimulus relationships, 132

- scaled test relationships, 126
- technique employed, 101
- template sounds, 109
- template technique, 106, 145
- peripheral hearing system, 147
- pitch, 25, 29, 30, 178
- presentation/environment, 25, 29, 30
- principal components analysis (PCA), 64, 130, 132, 200, 254
- quantification of effects, 85
- redundancy, 47, 172
- research techniques, 58
 - acoustical parameters, 62
 - perceptual distance, 61
 - semantic scales, 58
- research usage aspects, 46
- self-organising maps, 64
- semantic scales
 - research techniques, 58
 - spectral correlates, 76
- semantics, 86
- sonic character attribute (SCA), 38, 175, 216
- sound samples, description, 244
- specificities, 54
- spectral aspects
 - relative importance, 82
- spectral balance, 192
- spectral centroid, 192
- spectral contour, 69, 73
- spectral slope, 192
- steady state, 66, 72
- stimuli, 34, 40, 41
 - number of, 41
 - type of, 42
- stimuli, description, 244
- strong partials, 182, 187
- structure of perception, 54
- summary of chapters, 229
- synchrony, 71, 75
- synthetic sounds, 43
- technique in this research
 - overview, 65
- temporal evolution, 70, 74
- thesis components, 18
- timbral analogies, 45
- timbral aspects of spectrum, 66
- timbral features, 33
- timbral features of spectrum, 71
- timbral objects, 37, 242
- timbre
 - as a continuum, 37, 144
 - as absence of other qualities, 29
 - as complex composite, 33
 - as distinguishing quality, 28
 - as highly coupled complex, 36
 - as multidimensional entity, 31
 - as natural part of perception, 143
 - as non-independent set of attributes, 30
 - as non-linear construct, 37
 - as non-strictly defined part of perception, 34
- timbre definitions
 - overview, 24

research definitions, 38
specifics, 28

timbre space, 17, 19, 31, 38, 81, 85, 143,
216
defining factors, 34, 40

time-frequency resolution, 152, 157

time-varying characteristic codes, 198

traditional instruments, 43

transient aspects, 71

universality, 81, 172, 216, 223, 224

verbal description, *see* semantic scales

vision and audio, 51, 92