An exploration of timbral semantics related to the pipe organ

Alastair Christian Disley

Thesis submission in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Electronics

Department of Electronics
The University of York
United Kingdom

Supervisor: Professor David M. Howard
July 2004

Hypothesis

Within the timbre space of the pipe organ, some timbral adjectives can be shown to have common understanding and a consistent correlation with acoustic phenomena.

Abstract

Timbral semantics is the study of adjectives relating to timbre. This thesis takes existing research on timbral adjectives relating to single source sounds and applies it to the pipe organ, a complex multi-source musical instrument.

A substantial literature review suggested an analysis/synthesis methodology, and a preliminary experiment suggested that common understanding could be demonstrated for some adjectives. Further listening experiments gathered appropriate adjectives from English-speaking listeners and suggested that those adjectives could vary with age, geographic location and visual stimulus presented.

Seven timbral adjectives were selected for further study and used as rating scales in a series of listening experiments using both recorded and synthesised ensembles. Five of those words demonstrated common understanding, and several demonstrated consistent correlation with spectral features derived from acoustic analyses. The spectral and perceptual effect of a reverberant environment was also examined.

The hypothesis was proven, and the work was novel in both its scope and contribution to knowledge. Complex ensembles were explored both acoustically and perceptually, and the results have strong applicability outside of the context of the pipe organ.

Contents

Appendices

List of figures

List of tables

Acknowledgements

The author would like to acknowledge the following people who have helped in this research:

- Professor David Howard, for supervision and guidance throughout the programme of research.
- Tim Tozer and Peter Comerford (internal and external examiners respectively) for their helpful comments and suggestions in the PhD viva.
- Colleagues and staff in the Department of Electronics at the University of York and throughout the musical acoustics community who have assisted both in practical terms and with their discussion and suggestions.
- The University of York, for funding the research through a PhD studentship grant.
- Rob Stone and the Department of Psychology at the University of York, for use of their Web Response system in the listening tests.
- The many volunteer listeners from both the University of York and around the world, without whose willing efforts this research would have been impossible.
- The churches and organisations who permitted recordings of their organs to be made.
- The author's wife Helen and parents Ian and Linda, for their support through the genesis, realisation and writing up of this research.

Declaration

This thesis describes a programme of original research work by the author. This work is founded upon a literature review of work published by other researchers, and their work is fully referenced and acknowledged throughout the text. Portions of this work have been published as papers over the period of research, and these are referenced in the text. This work is otherwise entirely the original research of the author.

1. Introduction

This thesis presents the results of an investigation into the relationship between the adjectives used to describe pipe organ timbre and the physical properties of the sound being described. This chapter introduces the thesis topic and motivation, defines what "timbral semantics" is, introduces the pipe organ and builds a research strategy to guide the study described in later chapters.

1.1 Motivation and aims

This thesis aims to answer the following questions:

- What words do people use to describe the sound of the pipe organ?
- Do any of those words have a consistent understanding across multiple listeners?
- If so, are there any common auditory cues that can be correlated with those words across multiple listeners?

Acousticians often describe sound in terms of waveforms, frequencies, envelopes and spectra. But when most listeners are asked to describe a sound, or the difference between two sounds, they use adjectives such as "bright", "dull", "rich" or "nasal". Such words do not have a formally defined correlation with measurable acoustic phenomena, and have often been disregarded as being too imprecise or complex to be employed for describing sound. However, the fact that listeners instinctively use such terms for describing sound suggests that these words are an important means of sound classification to listeners and may give clues to high level auditory processing in the human brain. This suggests that they are worthy of further study even if they may not slip easily into established categories of timbral analysis.

The author's previous research (Disley, 1999 and 2000) has included work on the acoustics of pipe organs. When listeners were asked what made an organ good or bad for a particular role, as might be expected they used undefined timbral adjectives. Two of these, "blend" and "strength", were studied in greater detail, and within limitations appeared to have some common understanding. This suggests that further research, verifying these preliminary findings and expanding the scope both contextually and linguistically, is likely to prove fruitful.

1.2 Hypothesis

> *Within the timbre space of the pipe organ, some timbral adjectives can be shown to have common understanding and a consistent correlation with acoustic phenomena*

This hypothesis has been written to be as unambiguous as possible, but some of the terms and ideas contained within it require more detailed unpacking.

Since Ohm first related timbre to the harmonic spectrum in 1843, timbre has been difficult to define. The American National Standards Institute (ANSI) definition of timbre is:

> *'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."*

This is further qualified with this 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" (ANSI, section 12.9)*

This suggests that timbre is everything else that is not conveniently specified elsewhere, a "psychoacoustician's multidimensional wastebasket" (McAdams and Bregman, 1979).

An alternative definition is that given in the Little Oxford Dictionary:

> ***Timbre:*** *distinctive character of musical sound or voice apart from its pitch and intensity.*

The importance here is not so much the spectrum or waveform, a valid if mechanistic way of viewing timbre, but the characteristic quality it imparts and the notion of identification. Further definitions of timbre, including discussion of their weaknesses, can be found in Creasey (1998, p24-28). Timbre is essentially a perceptual quality, rather than something that can be easily measured or plotted on a simple scale. However, various attempts have been made to define the dimensions within which timbre can be quantified, and these are considered further in sections 2.2.1 and 2.2.3.

Timbre space is, therefore, a perceptual entity of multiple undefined dimensions in which all timbres can be found. Individual instruments or classes thereof (for example brass, strings, or saxophones) exist in subsets of that timbre space. Their common timbral characteristics often reduce the number of dimensions needed to adequately describe the differences between members of that group, and make the analysis methods described in section 2.2.1 practical.

Previous studies have defined limits or subsets of timbral space in order to improve the chances of getting meaningful results. These limits are usually a particular instrument or class thereof, with a specialised group of listeners from a single geographical and linguistic group (for example Kendall and Carterette, 1993 (1), Moravec et al, 2003, or Nykänen and Johansson, 2003). Some studies have ventured further, for example using individual samples of a number of different instruments and non-instrumental sounds (Creasey, 1998).

The timbre space of the pipe organ is interesting to investigate for a number of reasons in addition to those mentioned in section 1.1. The pipe organ is essentially a performer-driven timbral synthesiser, offering a number of different sounds that are acoustically blended together. The mechanisms of this and the uniquely controlled manner in which these sounds can be produced, in comparison to most acoustic instruments, are considered in section 1.4.

Timbral semantics is the study of words relating to timbre, and will be considered in more detail in section 1.3. The analysis techniques to explore the acoustic phenomena will be chosen or developed in light of previous research, and the nature of the acoustic correlation will be determined over the course of this research.

1.3 Introducing timbral semantics

The close relationship between our perception of timbre and the words we use to describe it has been introduced in section 1.1. Timbral semantics is the study of that relationship. Semantics is the "branch of linguistics concerned with meanings [of language]" (Little Oxford Dictionary), and it is important to note that it is not synonymous with "words" or "adjectives" but refers solely to their study.

Study of these words generates a whole range of questions in addition to those posed in section 1.1. For example:

- Do people use different words according to their particular musical taste, study, profession or ability?
- Do the words used by one person vary according to context?
- Is there any common understanding among certain groups of listeners?
- If so, how does this vary for other groups of listeners?

Some of those questions have been at least partially answered by recent studies, and these will be introduced in chapter two. This area has not been long studied for a number of reasons. It is often dismissed as inherently subjective, with the very real possibility of finding nothing of significance. There are a vast number of possible timbral adjectives in use. Moravec et al (2003) received 1964 different Czech timbral adjectives from a survey of 120 people. However, this experiment simply asked respondents to list all possible words which they might use to describe timbre. It is likely that if the words were asked for in a more indirect manner, a smaller subset of words in common use would emerge.

1.4 Introducing the pipe organ

The pipe organ is uniquely complex among musical instruments, giving a single player access to a multitude of timbral possibilities and dynamic levels unparalleled prior to the development of musical instruments incorporating electronics. Section 1.4.1 introduces the physical aspect of the pipe organ, including the way in which numerous pipes blend into a coherent ensemble, and section 1.4.2 describes why it is particularly suited to experimental use as proposed in this chapter. Alternative introductions to the pipe organ can be found in Thistlethwaite and Webber (editors, 1998), Anderson (1969), Sumner (1962) and Owen (1988).

1.4.1 The mechanics and sound production of the pipe organ

At the heart of the pipe organ are pipes. These are tubes, usually of circular or square cross-section, which produce sound when pressurised air is admitted to the lower end. The method by which they initiate and maintain the vibration of the enclosed air column that produces the sound varies according to the type of pipe.

Flue pipes are the most commonly found family of pipes, and work in a similar way to the recorder. These have a "mouth" at the base of the column, where a thin sheet of air crosses a gap before hitting a sharp edge (the upper lip). The oscillations of the air sheet either side of the lip interact with the body of air in the pipe to make the pipe "speak" (produce audible sound) at a given frequency. This frequency depends on the length of the air column and whether the pipe is open or closed ("stopped") at the upper end. In the latter case, the length of the standing wave within the pipe will be almost double, resulting in a lower pitched note.

The spectral characteristic of the pipe is affected by a number of different factors. Chief among these is the ratio of length to width of the pipe. Narrow pipes produce a spectrum with increased higher partials (Mercer, 1953, p379). These pipes are called "string" pipes, and in extreme cases require assistance at the mouth in order for oscillation to commence at all. Wide pipes in contrast have stronger fundamentals and reduced higher partials. These pipes are called "flute" pipes. Pipes between "string" and "flute" in both width and spectral content are called "principals", and make up the majority of flue pipes in most organs.

As well as the relative width of the pipe, other factors affect its speaking characteristics both in the time and frequency domains. The wind can be supplied at various pressures, and constricted at the base of the pipe or at the mouth. The mouth can vary in shape, dimensions and treatment, and the art of adjusting these parameters to gain a desired tonal quality is called "voicing". The overall construction of the pipe can also vary. It is unclear how the pipe material in and of itself affects pipe tone, but certain materials lend themselves to particular resonator shapes. The resonator does not need to have a fixed diameter, and some pipes have smaller pipes inserted at the top, which affects both tonal quality and resonant frequency. Tuning of flue pipes is usually achieved by adjusting the effective length of the resonating cavity.

As well as flue pipes, as described above, organs also have reed pipes. These produce sound by means of a single beating reed, in the manner of a clarinet. Unlike flue pipes, the resonator has only a secondary effect on the frequency of the pipe, which is primarily determined by the dimensions, material and stiffness of the reed itself. Tuning is achieved by adjusting the length of the reed that can freely vibrate. A reed without a resonator gives a very rich harmonic spectrum. Resonators modify this sound by their shape, either to imitate orchestral instruments or provide more generic organ reed tone. Longer resonators will typically emphasise fundamental and lower harmonics at the expense of higher harmonics, but the overall sound is still distinctly different from that of flue pipes. This is partially due to their more rapid onset transient and their more even distribution of harmonics.

The physical parameters affecting the different tones produced in a typical pipe organ have now been introduced. However, a pipe organ is not merely a collection of different sounds, but excels in the blending of those sounds into a single coherent ensemble. To understand this procedure, a brief overview of the mechanism of the pipe organ is necessary.

The player's interface with the pipe organ is via one or more keyboards and a pedalboard (a keyboard for the feet). Each key has a linkage, which may be mechanical, electrical or use some other technique, to a pallet. Pallets are air valves within the pressurised wind-chest upon which the pipes sit, and form one axis of a mechanical matrix. The other axis is that of the stops. Individual pipes are arranged into chromatic ranks of similar tone, with one pipe per note of the keyboard. Each rank is controlled by a "stop". If a player "draws" or activates a particular stop, the pipes of that stop will sound when the appropriate key is pressed. Drawing multiple stops will cause multiple pipes to sound.

Unlike the multiple sound sources of an orchestra, the many different pipes often speak at harmonics of the unison, altering the timbre of the sound as much as affecting the overall amplitude. This makes the pipe organ one of the earliest forms of timbral or harmonic synthesiser. The organist can combine the stops available in a number of ways to achieve timbres with certain desired characteristics. Manipulation of timbre through stops is often more intuitive than on synthesisers, as the organist is not dealing with abstract parameters such as those used in analogue or FM synthesis. Instead, the organist has a fixed set of complex sounds that can be combined in many different ways. For example, an eight stop organ (which would be considered small) has 255 possible combinations of stops. Many of those would not conventionally be considered musically useful, but that is still a large and intriguing palette. In a large organ, such as that of Doncaster Parish Church used later in this thesis, there are over ninety stops.

In a typical large pipe organ, there will be principal pipes speaking at the fundamental frequency and the second, third, fourth, sixth, eighth, twelfth and sixteenth harmonics. The higher harmonics tend to be grouped together onto a single stop called a mixture. As these mixtures go up the keyboard, ranks of the highest harmonics will tend to "break back" and duplicate lower harmonics, as very small pipes are not only difficult to manufacture and tune, but tend towards the limits of human hearing.

Those particular harmonics are chosen for their blending potential. As in the orchestra, "Octaves and fifths, in certain contexts (and depending on the tuning system) can easily be mistaken for single notes, because of their large coincidence of partials" (Sandell, 1991, p72). The pipe organ exploits this, and creates a cohesive blended ensemble by combining pipes speaking at many different pitches, all harmonically related to one another. Within pipe organs, the same can be said for other harmonic intervals to lesser extents. The major third (tuned pure, rather than tempered) is a component of many pipe organs (either as a Tierce rank, or as part of a Sesquialtera or Mixture with a 17th or 1 3/5' component).

This ability to select particular harmonics to synthesise a particular ensemble is taught in a basic way to organists. Harmonics are added from the bottom up to add perceived brightness to the ensemble, with absent harmonics only used to create timbres for solo effects.

The pitch of a particular rank of pipes is indicated by an approximate length in feet of the longest pipe in that rank. Thus a unison stop, for which middle C is approximately 256Hz, is known as an eight foot rank, as the open C pipe two octaves below middle C measures approximately eight feet in length. Although a stopped pipe only needs to be approximately four feet long to produce the same frequency, it is still known as an eight foot rank to avoid confusion.

The sound most commonly associated with a pipe organ is that of its principal ensemble. String pipes are rarely found at other than unison pitch. On larger organs, flute pipes can also be found at harmonic multiples to create synthesised solo sounds. These are sometimes grouped into a single stop called a Cornet. Larger organs also feature sub-unison pipes that speak an octave below the unison. Pedal organs are played with the feet, and are usually based on sub-unison pipes. Reed pipes are normally found at unison pitch, but can sometimes be found an octave above or below that.

To clarify the complex variety of possible timbres in the pipe organ, table 1.1 classifies the most common stop names according to family and pitch. The individual names vary according to small differences in construction and voicing. There are a particularly large number of names for flutes, both open and stopped, of 8' and 4' variety, most of which have been omitted to aid clarity.

Speaking length	Fundamental frequency at middle C	Principal	Stopped Flute	Open Flute	String	Reed
16 feet	128Hz	Double Diapason	Bourdon		Contra Gamba	Trombone, Fagotto
8 feet	256Hz	Principal or Open Diapason	Gedact or Stopped Diapason	Hohl Flute or Claribel Flute	Gamba, Viola or Salicional	Trumpet, Oboe or Clarinet
4 feet	512Hz	Principal or Octave	Stopped Flute	Open Flute	Salicet	Clarion
2 2/3 feet	768Hz	Twelfth		Nazard		
2 feet	1024Hz	Fifteenth or Superoctave		Piccolo or Flageolet		
1 3/5 feet	1280Hz	Seventeenth (rare)		Tierce		
1 1/3 feet	1536Hz			Larigot		

Table 1.1 – Commonly found pipe organ stops by pitch and family

Mixtures are usually made up of principal ranks speaking octaves, fifths and multiples thereof above the unison. Pitch designation is by a series of numbers, for example "Mixture IV 19.22.26.29". The "IV" indicates that there are four ranks of pipes. The other numbers refer to the number of diatonic notes above unison that the rank speaks. This nomenclature also applies in the case of the Twelfth and Fifteenth in table 1.1 above, both of which are commonly found stops. In the case of a fifteenth, the pitch of the lowest pipe is the same as the pitch of the fifteenth white note (c1) on the keyboard when a unison stop is drawn. The precise point at which mixtures break back is not usually indicated on the stop, and varies from one organ to another. This complexity makes comparative analysis of ensembles that include mixtures more difficult than those without.

The principal chorus consists of all the different pitches of principal stops on a particular manual being played together. A manual is another name for a keyboard, and usually there are two or more along with the pedal keyboard previously mentioned. The most complete principal chorus will usually be found on the "Great" manual. A "Choir" manual usually includes softer stops, and a "Swell" manual is enclosed in a box with adjustable louvres to provide a crude yet effective volume control.

The pipe organ relies on many different pitched pipes fusing together into a coherent ensemble, much as an orchestra relies on many different instruments combining to form a coherent ensemble. The blending of sounds from multiple pipes depends on a number of factors. The pipes have to be physically located within a reasonable distance of each other. What constitutes a reasonable distance varies with the size and nature of the acoustic environment. If the pipes are located sufficiently far apart, their perceived onsets may not be simultaneous, which is required for them to blend (Sandell, 1991, p7). The temporal acuity of humans can be as little as 2ms (Green, 1973). Onset differences of between 10 and 30ms, while they will not be perceived directly by the listener, will suggest the presence of two tones rather than one (Rasch 1978).

Their locations may also appear to be different to the human auditory spatial positioning mechanisms. The case or immediate environment of the pipes helps acoustic blending, and some organ cases have been specifically designed to blend and focus sound. However, some organs without such cases still produce a coherent ensemble, and the change that a constrictive case or swell box has on the overall sound is not always desirable. Blend in pipe organs is further considered in section 2.2.

Over the past two centuries, the specification of a good timbre of a particular class of stop or of a complete ensemble has changed many times. This is partly due to changes in the technologies available, and partly due the roles the organ has been asked to perform. Many instruments are now thought to have been altered to their detriment in line with taste at the time. Bicknell (1996) provides an English overview of this.
The motivation behind the changing tastes for different tonal qualities is outside the scope of this introduction, but can be found in the pipe organ texts referred to at the beginning of section 1.4. However, the fact that tastes have changed is important when attempting to make judgements of musical tone. Churcher says:

> *To the individual listener, a good timbre of a particular class is simply the one that he favours. In general, a good timbre is one favoured by a majority of competent listeners. Whether this is ascertained by test or simply by survival in organ-building practice it is, of course, this criterion which prevails in the end. It is better to please than displease the majority."* (Churcher, 1970, p125-6):

1.4.2 The pipe organ as a tool for experimental use

Pipe organs are some of the most well documented acoustic instruments, and it is possible to take controlled samples from them in a way impossible with most other acoustic instruments. When recording acoustic instruments, one problem is that two similarly pitched and played notes will never be identical in terms of their spectrum or waveform (Freedman, 1967). The pipe organ gets as near to that as possible, because the way in which the player triggers the note has no effect on the steady-state spectrum. If it the pipe organ used for recordings is a mechanical action instrument, the player may be able to alter the transient portions. However, this is only possible to a small degree when compared with orchestral instruments. This predictability makes it easy to take controlled and repeatable samples from a pipe organ.

Hall (1993) describes this in the following manner:

> *"The pipe organ offers the opportunity to conduct psychoacoustic experiments in which the sound of a natural instrument can be perfectly steady and reproducible.*

The pipe organ also offers a number of possibilities for examining different timbres. Any reasonably sized instrument has the potential to produce a number of different timbres by means of the traditional stop control. Comparison of different instruments may determine the role of different tonal qualities in achieving certain timbral effects. The pipe organ has also been reproduced in electronic form by several specialised synthesis systems. The Department of Electronics at the University of York owns one of these, developed by Bradford University. There may be circumstances where the use of a controlled synthesised sound source is preferable to an acoustic source. The synthesis technology also permits fine adjustment to the sound in a way that would be impractical to carry out with a pipe organ. The Bradford synthesis system will be introduced in more detail in section 2.3

1.5 Development of initial research strategy

It is useful at this point to refer back to the hypothesis, here to examine how the principles behind it can best be explored and tested.

> *Within the timbre space of the pipe organ, some timbral semantics can be shown to have a consistent correlation with acoustic phenomena*

What is the timbre space of the pipe organ, for the purposes of this research? Is it every possible sound or combination of sounds that any pipe organ can produce? Or is it the limited subset thereof that has been declared useful by organists over the centuries? While restricting the timbre space further at this point may seem premature, it is sensible to use the most commonly used subset of potential sounds as a starting point. This would mean that the results obtained would have maximum applicability to the most commonly found organ ensembles. By beginning to explore the timbre space of the pipe organ with ensembles familiar to listeners, the chance of discovering common understanding is also increased, as such agreement is unlikely to be found with unfamiliar ensembles.

For similar reasons it is also important to present the ensembles in a musically appropriate context. Most research to date on the pipe organ (see section 2.1 for a summary) has been analysis of recordings of single pipes. While this is entirely appropriate for detailed analysis of acoustical properties, to use single samples in listener based studies unnecesarily restricts the scope to a limited selection of sounds that are not normally heard on their own in a musical context. The pipe organ needs to be considered in terms of its ensemble as well as the constituent parts thereof.

Most of the single samples used in research are also anechoic. While this is again appropriate for certain analyses, commonly received wisdom of organ builders states that "the room is the most important stop on the organ". Lady Susi Jeans said "I think that favourable acoustics for organ tone can do more for an organ than the organ builder himself" (quoted in Churcher, 1965). Voicing of a pipe organ to match the particular resonant frequencies and idiosyncracies of the acoustic environment in which it is to be placed is an important and lengthy process in the creation of a blended ensemble. To ignore the acoustic environment when considering a particular pipe organ as a whole is therefore inappropriate. Stops that may seem unevenly voiced at close proximity may prove perfectly balanced from the usual listener position in the room.

Goad and Keefe (1992) have demonstrated that variations in listener position produce variations in perceived timbre for orchestral instruments. Syrovy et al. (2003) have begun to discover similar effects with pipe organs. It is therefore important to record the pipe organs for listener rating from the position in which they were meant to be heard. Goad and Keefe also compared melodies with isolated tones, and discovered no difference in listener discrimination of timbre. This suggests that presenting listeners with examples in an appropriate musical context, as suggested above, would be no less rigorous an experimental approach. It is important to note that what is being proposed here is not that the listeners should be played sounds and asked how the sounds make them feel, which would be an emotional experiment in which the style of music would be very significant. Instead it is asking them to describe the sound itself.

The nature of that musical context may vary. Disley (2000) discovered a difference in rating of pipe organ ensembles depending on whether the musical examples specifically used polyphony (such as counterpoint), or whether it was more chordal in nature (such as hymnody).

Within the context of commonly available pipe organ ensembles, there are several options available for verification of any theories developed, as presented in section 1.4.2. Artificial synthesis of ensembles with controlled adjustment of parameters could be used, or real pipe organ ensembles could either be compared and analysed or used as synthesisers in their own right through use of different stop combinations.

1.5.1 Existing research on timbral semantics

One potential avenue for exploration as part of an initial research strategy is the plethora of existing theories of timbral semantics that exists. While this is not a well-studied area, there are a number of previous studies, which are presented in detail in section 2.2 of the literature review. An examination of what is already known is an essential part of any research strategy. Some of the documents referred to in chapter two were published after the commencement of this piece of work, and therefore may supersede or contradict assumptions made in earlier experiments described in this thesis.

Suitable criteria could include the Grey attributes, those identified in addition in Sandell, and those summarised in Creasey (1998, pp. 76-80). Adjectives that stand out as particularly suitable for exploration in this timbre space include brightness, clarity, dullness, hardness, weight (heaviness or lightness), nasality, richness, reediness, thinness, and warmth. Which of the words used would be useful to rate pipe organs?

However, these are terms gathered second or third hand from others' experiments. The timbre spaces they have been gathered from are mostly not related to pipe organs, and indeed some semantics will have been postulated rather than formally gathered. This piece of research would be given a firmer base by gathering timbral semantics specifically for the timbral space being investigated. This could be achieved by asking listeners to describe the difference between recordings of different pipe organ timbres or asking them to describe a timbre to someone who had not heard it.

These adjectives could then be compared with acoustic analyses of the timbres played to the subjects, to see if correlation existed. These analyses could include some of the methods detailed in Grey and Sandell, although many of those analyses are simply too complex for the combinations of multiple pipes dealt with here. If correlation exists, and the research on "blend" and "strength" in Disley (1999 and 2000) suggests that they do, then they can be tested by the synthesis of realistic examples which should (or should not) exhibit those properties. These new theories can then be verified in listening tests. A good starting point for research would be to apply this synthesis method to the results found in Disley to see whether they can be verified.

It may be appropriate to establish by some form of analysis (e.g. MDS) what the significant axes of the timbre spaces explored are, but this is not a primary goal of this study. The multiple-note samples proposed in section 1.5 may make such analyses difficult to obtain, and their relevance is questionable. A more achievable aim is to prove that that common understanding and perception exists, and then attempt to determine the triggers behind that as time permits. It may be possible to assess which of the timbral axes from previous work might be appropriate for multiple notes in a natural environment.

1.5.2 Linguistic considerations

Timbral descriptors inevitably differ between subjects based on nationality, language and even dialect. It is important to consider how to deal with this in an experimental regime. Definitions of the words could be provided, or an attempt could be made to select subjects with a common linguistic background. The problem with providing definitions is that it risks defining the term to match the description already ascribed to it, thus falsely encouraging the results to tend towards confirmation of theories. One possible way around that would be to work from separately developed definitions, gathered from subjects at the same time as the adjectives, but this is still problematic.

The problem with using subjects with a common linguistic background is that it limits the scope of the experiment and the uses of its results to some extent by requiring a carefully selected subject panel. Sandell addressed this question of language, and claimed his results were consistent for Western musical listeners (Sandell, 1990, p246). However, he did not specify where his listeners were from other than to say they were students or teachers at his university (Northwestern University, USA). It is possible that results could vary across borders on several scales, but provided that data is gathered about the linguistic background of subjects, there is no reason why a larger subject group could not be used, with later breakdown of results into specific linguistic groups.

1.5.3 Sampling and synthesis considerations

If synthesis of sound examples as proposed in section 1.5 is to remain within the timbre space of the pipe organ, it must not sound artificial. How far should the pursuit of realism go? It is clearly important to include all factors that are involved in the perception of the particular timbre, but that need not stretch as far as including extraneous sources prevalent in natural acoustic environments such as traffic noise or more organ-based distractions such as general wind supply noise. It should be noted that at least one recent large electronic installation (Marshall and Ogletree, Trinity

Church, Wall Street, New York) simulates both wind supply and swell shutter noise, but this is irrelevant in specific consideration of timbre.

The sound must be as realistic as possible so that it is applicable to real timbre spaces. One major reason for sounds to be judged electronic is harmonicity (Strang, 1969). Certainly any inharmonic components should be considered in both analysis and synthesis. Current synthesis technology is introduced in section 2.3 to facilitate the generation of realistic pipe organ timbres.

Another major cue to the artificial nature of sound is the reverberation. Its significance has been discussed in section 1.5. Within the Department of Electronics exists a Yamaha S-REV 1 convolution reverberation unit, which takes measurements of actual room acoustics and then recreates them. Recordings played back through such a unit can be placed in the virtual room. This makes it possible to place an organ that has been recorded at close range in a given room, and place a listener elsewhere so that the full benefit of the room acoustics can be felt.

When recording the original samples of pipe organs, the procedures outlined in Horning (1998) will be followed. When recording the ensemble at close range to permit easier resynthesis, it is important to note that sound radiates from both openings in an open flue pipe. Miklós and Angster (2000) suggest placing microphones at 3-5cm from both ends to avoid phase problems. They also report a dip in the 6th harmonic at the mouth when compared with the end. An alternative to recording a standard stereo signal is that proposed by Goad and Keefe (1992, p45) who suggest that for timbral analysis, a dummy head binaural device may be of more use. However, this would be impractical for use within an organ, and would require listeners to use in-ear headphones. Ideally, casts of each listener's ears should be attached to the dummy head when recording, but this is obviously impractical in the context of this experiment.

1.5.4 Experimental procedure

The nature of the testing medium needs careful consideration. In Disley (1999 and 2000), the author experimented with different Internet based techniques. Ideally subjects would be co-located, but players and regular listeners of the pipe organ are not a common subject group, and to gather a significantly sized group together might prove expensive and impractical.

Presenting sound recordings to subjects over the Internet has a number of potential problems. Experimentally, each subject's headphones may have different frequency responses, so in any precise experiment it is essential that samples be presented comparatively. This compensates for the inevitable differences in experimental environments. Practically, subjects must download large audio files, as the effect of compression algorithms such as the MPEG audio standards on psychoacoustic testing is not currently known. With the advent of broadband and other cheaper ways of accessing the Internet, downloading does not take as long as it once did, but the time spent downloading must still be minimised in order to continue to engage subjects with the experiment.

An alternative method of remote subject testing explored in Disley (2000) was to provide the subjects with pre-recorded CDs. However, comparison tests and repeating of tracks made experimental error more likely as subjects had to operate the controls of the CD player themselves. Any such future subject controlled tests should be sequential in operation, preferably with clear spoken instructions at the start of each track to reduce the possibility of listening to the wrong example.

Both remote experimental techniques offer a large increase in subject group size at the expense of potential experimental accuracy. It is important that any such remote experiments take care to minimise this risk, and are accompanied by local experiments with smaller groups of subjects to detect any difference between the two groups. One simple way to achieve this is to present the local group with the same Internet based test on a single computer with high-quality headphones in a controlled environment. The test could be run from the local hard disk, thereby avoiding most of the issues of Internet testing for those subjects while not altering the testing environment from that of the remote subjects. Guidelines for setting up such experiments to minimise variables and ensure listener comfort are given in Levitin (in Cook (1999), particularly pp. 312-313), Mursell (1964), Davies (1978) and Von Békésy (1960).

1.6 Summary of research strategy and conclusions

The initial motive of the author's previous work was to move away from ill-defined timbral adjectives to a better-defined analysis of the pipe organ. However, as a range of acoustic analysis methods (detailed in chapter two) have been developed, the sound itself is well described in terms of frequencies, envelopes and so on. What is less well understood is how that is then perceived by the listener, and what methods the listener uses to classify and describe that sound. Churcher (1970, p116) says:

> *"So study of the connection between sound spectrum and timbre has the handicap that while the former can be explicitly stated, the latter can be indicated only in broad terms."*

However, the success of experiments by the author and others described in section 1.5.1 and in greater detail in section 2.2 suggests that timbre can be indicated in more precise terms. Section 1.5 sets out the case for further research in this area and considers a number of appropriate experimental techniques.

The proposed research will commence by study of existing timbral semantic work, and look to verify those findings within the timbre space of the pipe organ. Analysis of the results of those initial experiments should then fine-tune the next stage of work, which will look to gather timbral adjectives from listeners in a non-biased manner. The most common of those adjectives will then be used as comparative ratings scales in listening tests. The tests will determine whether there is common understanding of any of those words, and acoustic analysis will be used to look for the causes of any common understanding.

When considering a research project, it is important to examine whether the proposed topic is worth looking at. Sandell (1991) uses a selection of criteria that are useful to consider when deciding whether a topic is worthy of further investigation. They have been altered slightly here, but only to expand their meaning outside of his particular experiment.

- Definability: is there a clearly defined objective?
  Yes. The hypothesis in section 1.2 states clearly what the objective is for this investigation.

- Demonstrability: Are there examples of successful or unsuccessful instances of the sonic effect?
  Yes, several examples have been described from the author's previous work.

- Continuity: Can the effect be experienced to various degrees (i.e. along a continuum as opposed to it being an "either-or" phenomenon)?
  Yes, previous research has demonstrated this for several timbral adjectives.

- Relevancy: Is the topic relevant?
  Yes, the topic is relevant to both psychophysical and acoustic research, developing existing basic understanding of complex phenomena.

To these the question should be added for doctoral work: "Is the work to some degree novel?" There are several things that are novel about this approach. Firstly, it moves from the controlled, artificial timbre space of Grey or the limited set of single samples used by others into a realistic timbre space. Secondly, it seeks to generate its own relevant set of adjectives instead of starting with a presupposed set. Thirdly, it looks to develop theories in a timbral space that has not previously been explored. The exploratory nature of much of this study is justified on the basis that this entire area of research is exploratory. Future experiments can have a more specific goal in light of the knowledge gained from this exploratory work.

The result of this thesis should confirm or deny whether existing tentative theories hold true in this timbre space, and create new theories with some degree of verification from listener tests. So in its scope, approach and hypothesis, this work is novel and will be a significant contribution to knowledge in this area.

2. Literature Review

The research introduced in chapter one falls into two main areas: acoustic analysis of the pipe organ (section 2.1), and studies of timbral semantics (section 2.2). Literature relating to the workings of the proposed synthesis system is also reviewed in section 2.3.

Familiarity with the basic workings and principles of the pipe organ as described in section 1.4.1 is assumed throughout this thesis. A good alternative text for gaining this information is Thistlethwaite and Webber (editors, 1998), and other suitable texts include Anderson (1969), Sumner (1962) and Owen (1988).

The perception of timbre is a cognitive analysis of information gathered through the human ear. An introduction to the human hearing system by Max Matthews can be found chapters one and two of Cook (1999). An introduction to the concept of timbre by the same author can be found in chapter seven of Cook (1999).

2.1 Acoustic Analysis of the Pipe Organ

The sound of the pipe organ has been analysed by techniques such as Fourier Analysis for as long as these techniques have been available. These techniques generally fall into two areas: analysis of the steady state and transient parts of the waveform. Much previous investigation into pipe organ tone has been conjectural, and that conducted before 1970 is summarised in Churcher (1970).

2.1.1 Steady State Analysis

One of the first people to use steady state analysis was Boner, (quoted in Bonavia-Hunt, 1939), who analysed individual pipes and ensembles to prove that the Hammond organ was not an accurate reproduction of the real thing. Frobenius and Ingerslev (1947) used Boner's somewhat cumbersome but accurate analysis of harmonic strength. They carefully (with the limited technology available at the time) analysed several pipes and plotted their harmonic spectra, and how they varied with the position of the microphone along the body of the pipe. Bergweiler et al. (2003) have done similar measurements using piezo-electric film, although their research is at a preliminary stage and while demonstrating successful methodology has yet to produce significant results.

Many people investigating the speech of flue organ pipes, including examination of the end-correction (what effect on pitch the open end of the flue pipe has), have used similar Fourier-based methods of steady state analysis. The number and scope of experiments looking at flue organ pipes is such that much of the information gathered is outside the bounds of that useful to organ-builders. Little of it is directly relevant to this thesis, so interested readers are directed to the summary in Miklós and Angster (2000, p612).

The author has developed previous methods of steady-state frequency analysis that plot the amplitude of each harmonic (Disley, 1999, ch.7, pp. 1-2). A third dimension was added so that measurements taken across the keyboard could be plotted on a single three-dimensional graph. The resulting 3D surface was termed an "Acoustic Signature". This provides a useful visual means of appreciating and comparing the steady-state harmonic spectrum of an ensemble, normalised by harmonic (Disley, 1999). This method was later partly automated (Disley, 2000, pp. 18-20).

Syrový et al. (2001) have developed a method similar to the author's, normalised by harmonics, to analyse the harmonic spectrum of a number of Czech baroque organs. They did not add a third dimension, so the results are presented in four separate graphs. To fit within journal space constraints, their harmonics scale only includes the harmonics that were present. This makes comparison of the graphs difficult, as each has a unique scale. One advantage of their method is that several organs can be plotted on a single graph for each note, but this can get confusing with many organs, particularly where the graphs are small and not reproduced in colour.

The acoustic signature and other similar methods of steady state harmonic analysis by their nature do not include information on the transient portions of the sound. Nor do the methods that present only the harmonics take account of possibly significant inharmonic components. Miklós and Angster (2000) suggest that there are significant inharmonic components in pipe organs, which may be important to consider in future analyses. These eigenresonances (resonances of the pipe body not directly associated with its speaking length) are often near harmonics, and can be detected by applying white noise excitation to the pipe and measuring the response. The resonances may also be what some people informally refer to as the formants of the pipe. Relative to the harmonics of the pipe in speech, their frequencies are slightly stretched. This stretching is particularly noticeable in wider scaled pipes, and least noticeable in tapered pipes.

The methods summarised in this section have been applied to recordings made at various differences from the pipes, with only the piezo-electric measurements of Bergweiler et al. requiring physical contact with the pipe being measured. Disley (1999 and 2000) has successfully applied the acoustic signature method to recordings made both in close proximity to the pipes and in a listener position some distance from the pipes. The author also used the method to analyse the steady state of a combination of pipes.

An integral part of the construction of a pipe organ is voicing each pipe in situ for the environment into which it must speak. A pipe's spectrum will typically be affected by both its immediate surroundings (the resonance or damping of the organ case and pipes) and also the acoustics of the room in which it is placed. Syrovy et al. (2003) have begun to verify this by comparing the spectrum of a number of pipes when recorded at three listener positions. These positions are within the organ (where the pipe would be voiced), at the console (the performer's listening position), and in the room (an ideal listener position). As expected by received wisdom, the spectra of single pipes varied according to where they were recorded from, and further research is planned by Syrovy et al. to provide sufficient examples that generalisations can be made of the effect of listener position on the spectrum heard. This also agrees with more the general understanding that "spectral balance as well as overall amplitude provides cues to the intensity versus distance of a source" (Shepard, p26, in Cook, 1999), which is covered further in section 2.2.3.

Goad and Keefe (1992) also examined whether timbre was altered by the acoustic environment it was in, such as from one concert hall to another, or indeed one seat location to another. They discovered that variations in the listener position did produce variations in timbre. When the listener results for melodies and isolated tones were compared, there was no difference in listener discrimination of timbre. This may not be the case when more than one note is played at once.

2.1.2 Acoustic Analysis of Pipe Organ Transients

Transients have long been considered as important in pipe organs (e.g. Kitching, 1987), but little evidence had been provided for this, and there is little literature on this subject area.

The author has modelled a test after Berger's orchestral instrument transient significance test (Berger, 1965). The results indicated that while the importance of transients varies with the stop chosen, transients always improve the identification of one stop over another (Disley, 1999, chapter five). The relatively low success rate of identification (37.2% with transients) was explained by the similarity between many of the samples. Taking this into consideration by combining the results for similar stops improved the result to 58.4% correct listener judgements, compared to 50.4% without transients. This verified both Taylor (1976) and Kitching's assertions that the onset transient portions of individual pipe organ stops were significant in the identification of the stops. The author plans further tests in this area.

Saldanha and Corso (1964) also conducted a test similar to Berger's, asking people to identify instruments. Their result of 40% correct listener judgements compares with the similar initial results of the author. They do not seem to have made any allowance in the test for the fact that listeners may have misidentified similar instruments, such as a violin for a cello. Practice helped their listeners to improve their judgement.

2.2 Timbral semantics and the pipe organ

Certain tenets of timbral study are applicable across many musical situations. Donald Hall (1993) states that "The pipe organ offers the opportunity to conduct psychoacoustic experiments in which the sound of a natural instrument can be perfectly steady and reproducible." (p417) Timbral semantics have been inherent to the description of pipe organ timbre for many centuries. Most writers use adjectives frequently and freely, but without formal definition. Within pipe organs, there is a tendency to describe pipes by their family: string, flute, diapason or reed, each of which has a distinctive timbral characteristic (see section 1.4.1).

Specific theories relating pipe organs to timbral semantics are rare. Mercer (1953) suggests that the perception of blend within a combination of stops would only be upheld where higher pitched stops reinforced the harmonics already existing in lower pitched pipes. This would suggest that a stopped 8' flute, with its almost entirely odd spectrum of harmonics, would never blend with a 4' stop (which speaks at the octave, thus filling in the missing harmonics). This is not borne out in practice.

The author investigated "blend" related to pipe organs (Disley, 1999 and 2000). Initially listeners were not given a definition of blend. McVicker (1987, p8) describes well-blended pipe organ ensembles as ones created by the organ-builder carefully scaling the pipes to match the room (see section 2.1.1). Disley (1999) concluded that blend was associated with a smooth acoustic signature – one that has no prominent peaks or troughs but smoothly tailed off from the low to high frequencies. This was verified by further testing (Disley, 2000) with the caveat that it only held true for an ensemble of principal pipes and not necessarily ensembles in which reed stops were present.

The author's works also investigated "strength", which was similarly undefined initially. Disley (1999) concluded with many caveats that strength was related to a firm 8' and 4'. The strength theory was refined in Disley (2000) to: "for an organ to be perceived as strong, it should have both solid unison (or sub-unison) foundations, combined with a rich harmonic development reinforcing the foundation's harmonics." Strength increased dramatically for reed stops (Disley, 2000, p47).

Originally, both works were an attempt to determine what made a good organ. Disley (2000) concluded that it was impossible and not a little naïve to attempt to define a good organ in one sentence or one theory. This would ignore the reality that different organs are good for different things. In Disley (2000) musical examples were played in two different styles: a contrapuntal work (a Bach fugue) and a hymn tune. People's preference ratings for the selection of organs were different depending on what was played on them.

Rioux (2000) has done some research linking semantics commonly used in describing pipe organs by organ-builders and organists with acoustic phenomena. By discussing results and analyses with an organ builder and applying multidimensional scaling to the result, he deduced that "chiff" was affected by the speed of the transient and one other unidentified factor. This use of pipe organ specific semantics – words that are known to have a particular meaning in the description of pipe organ tone – could be a good starting point for further consideration. Potential candidates include "reediness", "stringiness", "flutiness" and so on. From personal conversation with organ builders, the author has discovered that most of the timbral semantics used by organ builders usually refer to specific physical alterations to the pipes and lack consistent definition, particularly from one builder to another. The range of words used is also surprisingly small.

2.2.1 Multidimensional Scaling

The science of correlating the physical properties of a sound to the perceptual attributes they give rise to is known as auditory psychophysics. The traditional methodology associated with this science is to measure the physical properties of the sound and the behavioural response. Then various methods, such as best fit and regression, are used to link the two. With regression a polynomial is generated, so within certain strict boundaries, an equation can be created giving a mathematical definition of an adjective in terms of acoustic phenomena.

Multidimensional scaling, abbreviated to MDS, is a method of taking a variety of timbres and placing them in a timbre space with the least number of dimensions possible. Efforts are then made to correlate various acoustic phenomena with the dimensions, with varying degrees of success – it is rare that the primary (most significant) dimensions can not be attributed to something with a reasonable degree of accuracy. A similar method of analysis is Principal Component Analysis, or PCA.

A seminal work on the understanding of timbre is that described in Grey (1977). Grey selected and synthesised a number of orchestral instruments, then attempted to classify them by playing them singly to subjects. Grey's hypothesis was that for an N-dimensional MDS solution, the listeners' timbral judgements were driven by N interacting factors.

He was able to reduce his set of dimensional scales down to three by looking at those that were statistically significant. These were, in order of significance, spectral energy distribution, synchronicity of the higher harmonics, and the presence of low-level, high frequency energy in the initial attack segment. Such results can only be valid within the timbre space used in the experiment. In this case, that space consisted of single notes of a small set of synthesised orchestral instruments, but Grey's work was a significant step forwards in knowledge.

The timbre space of Grey's tones was relatively small (Grey and Moorer, 1977), and many of the listeners found it hard to distinguish between the pairs of tones. In later experiments, Grey and Gordon (1978) verified Grey's spectral axis.

Sandell (1991, p184) examined Grey's dimensions in more detail. He confirmed that the Y-axis correlated with the spectral centroid (the centre point of spectral energy distribution). However, he discovered that the Z-axis did not correlate with precedent attack noise, but instead with the duration of perceived attack time or DPAT. It also correlated with harmonic envelope synchrony. The X-axis had correlations in Sandell's thesis with all three measures of synchrony: peak synchrony, harmonic envelope synchrony and on-off synchrony. The harmonic envelope correlation appeared to be for all frequencies, not just the upper ones as Grey had stated. Sandell concluded that this dimension was less distinct than the other two, as the PAT (Perceived Attack Time) and recognisability also correlated with the X-axis.

One failing of MDS is that each sample must be compared with every other sample. This means that for each extra sample introduced, the number of tests that must be carried out is doubled. Scavone et al. (2001) have developed a method to combat this, which involves the user being able to move the sounds around their own, undefined, two-dimensional timbre space. This space is represented on a computer screen, with subjects using familiar drag and drop techniques to group the timbres according to similarity or difference. Colours can also be used to aid this, or indeed to create a third dimension of sorts. This greatly reduces the strain on the user in terms of the time taken to undertake the task, but does make subsequent analysis potentially more difficult. One interesting note is that their subjects commented that they found that method much more engaging than a traditional pair-wise methodology.

However, all of the studies mentioned thus far in this section have tended towards a solution with a limited number of timbral dimensions. This is an inherent tendency in MDS, and one that has been questioned by a number of authors (e.g. Kostek (1999), summarised in Kostek et al, 2003). This limited number of dimensions is attractive to authors, as it is easier to comprehend two or three dimensions in a visual form.

Creasey (1998) notes that the tendency of previous studies to dwell on three or four dimensions does not exclude the possible existence of other dimensions, which may be important (p48). He cites the 3D solution in Plomp (1976) that accounts for 90% of differences, still leaving 10% undetermined in a very small set of sounds (Creasey, 1998, p49). Padgham (1986) suggests a two-dimensional solution based on visual techniques, and Pries (1984) suggested more generally that timbre had one significant dimension to which all others were nuances.

Bregman (1990) suggests that the temptation of a low number of dimensions is probably partially due to the fact that all available colours can be described in three dimensions (p122). However, colour itself is not the entirety of visual experience. The setting, surface texture and light conditions affect the colours presented to our retinas, and subsequent visual analysis determines the real colour of the object, allowing for such effects as coloured light. Our auditory processes may do something similar.

PCA emphasises that certain components may be significant for a given circumstance, but the axis titles generated by this method vary, indicating that certain timbral qualities may vary in importance depending on the situation. Bregman (1990) suggests that different sample sets will probably have different major axes of timbre space. When subjects are asked to judge differences for a given set of stimuli, they will tend to pick up on the two to four most salient ones (p124). So the three or four dimensions of an analysis may reflect those which are most significant for a particular sample or set thereof, but those significant dimensions would not necessarily be the same in a different test with different examples (Creasey, 1998, p50). Creasey also concedes that the major axes of timbre can be extracted from a small set of sounds (p42). PCA is less limiting in this respect than MDS, as description of all factors is not attempted. MDS also does not suggest labels for the axes, and the problems in Grey's work identified by Sandell indicate the difficulty of reliably recognising an axis.

2.2.2 Blend

Blend is a timbral adjective that has more general application than most. Sandell (1991) investigated the concept of orchestral blend in detail. His motivation was to reform the theories of orchestration, so that they were based on auditory perception and thus were equally applicable to new musical instruments, such as synthesised sounds. Where Grey's experiment was about the similarities between instruments, Sandell's started with the same set of tones as Grey, but looked at their concurrent properties.
He defined blend as:

> *"The tendency for concurrently sounding timbres to fuse into a single timbre."*
> (Sandell, 1991, p7)

This expands upon the more blanket dictionary definition of blend as to "mingle intimately; become one" (Little Oxford Dictionary, edited by J. Swannell, 1986) without disagreeing with it. He goes on to say "Interpreted in musical terms, blended combinations would be those in which the distinctiveness or individuality of the constituent instruments is subordinated to obtaining an overall, uniform timbral quality." (Sandell, 1991, p40)

Sandell discovered that the primary influences on blend were the centroid position on the frequency axis, and the duration of the onset for tones. When the tones were in unison, the most important factor in creating blend was the average centroid: the lower the better. He repeated the experiment with tones separates by an interval of a minor third. Then the most important factor in creating blend was that the spectral centroids of the two instruments were close to one another in frequency. Another significant factor identified from those considered was the correlation of amplitude and centroid envelopes. Blend increased as temporal patterns rose and fell in synchrony. Similarity in the overall amount of fundamental frequency perturbation was also important, as there was decreased blend with increasing jitter from both tones.

Sandell's formula for the calculation of spectral centroids is:

$$f_c = \frac{\sum_{n=1}^{h} f_n a_n}{\sum_{n=1}^{h} a_n}$$

where f_c is the spectral centroid, h is the number of harmonics, f_n is the frequency of harmonic n, and a_n is the linear amplitude of harmonic n. This formula is suitable for steady state waveforms that are periodic and harmonic in nature, over a fixed time period. Sandell gives a variety of formulae to calculate the centroid taking account of variable lengths of time, and tones with temporarily variable amplitude and frequency functions for different harmonics (Sandell, 1991, p155).

His discoveries relating to blend have parallels in other research. Goodwin (1989) discovered that when solo singers were compared to singers who were part of a choir, choral singers altered their voices to improve blend. They reduced both the number and strength of upper partials, and emphasised the first vocal formant at the expense of the second and higher formants.

Dannenbring and Bregman (1978) demonstrated that a darker spectrum (which they defined as one without prominent high frequency components) aids the fusion of harmonically related partials into single tones. According to their research, the best scenario for fusion would be a spectrum where low frequencies were of greater amplitude than high frequencies. This should be accompanied by a smooth progression from one to the other, with no prominent high harmonics or absent lower harmonics (see figure 2.1).

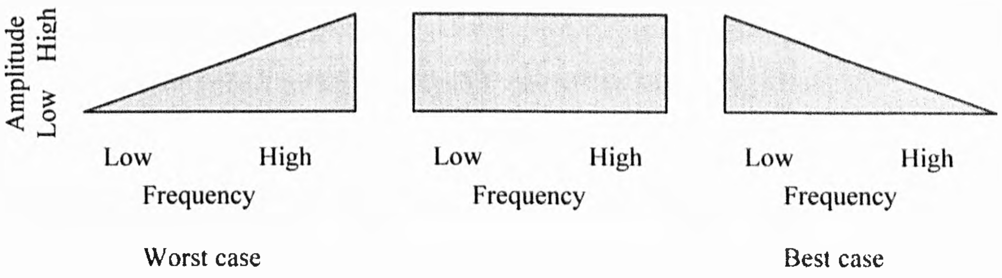


Figure 2.1 – Best and worse case frequency spectra for perceived fusion, after Dannenbring and Bregman (1978, p372)

This is analogous to the smooth acoustic signature proposed for blend in pipe organs by Disley (1999), described in section 2.1.1. Organs rated best for blend had an acoustic signature that looked similar to the best case in figure 2.1 with no harmonics either prominent or absent.

Sandell also described a number of different acoustic analyses that he undertook to determine if any of them correlated with any of the dimensions of his MDS results. One such method was a measure of dissonance, for an aggregate of harmonics from one or more complex tones, after Hutchinson and Knopoff (1978).

Sandell considered three types of synchrony within a tone (Sandell, 1991, p168). Harmonic synchrony is where the amplitudes of the harmonics follow a common path. Peak synchrony is where the peaks (in more than one tone) occur at the same time. Onset and offset synchrony is where harmonics start and stop at the same time. This is similar to the common fate theory described in Bregman (1990, pp. 248-292), where auditory events that share a common trajectory in some dimension are perceived to be related. Sandell then looked at pitch deviation, which considers the average pitch, and the difference of trajectory between the pitches of two tones, as described in Gold and Rabiner (1969).

In total, Sandell considered eleven possibly correlated factors. These were:

> the spectral centroid, acoustic dissonance, precedent noise, duration of PAT, amplitude scaling, harmonic envelope synchrony, peak synchrony, onset/offset synchrony, harmonicity and pitch deviation and recognition.

Correlations Sandell discovered suggested that a large proportion of dissonance could be attributed to inharmonic activity in the attack portion, as it was correlated to DPAT and precedent noise.

One interesting fact Sandell discovered was that of Grey's tones, the reeds blended worse than other timbres (Sandell, 1991, p243). This concurs with the author's research in which pipe organ reed stops blended worse than flue ensembles (Disley 2000). Sandell criticises some aspects of his own approach, such as the regression method, which generated many possible factors for the two dimensions discovered in his MDS for blend. He ended up disregarding many as insignificant, and came up with five factors, which were, in order of the most important:

> Centroid sums, composite dissonance, pitch deviation sums, tristimulus representation (upper harmonics only) and DPAT absolute differences.

The tristimulus representation is considered in section 2.2.4. Sandell concluded that seven of his eight subjects showed a consensus of what blend was. He claimed this was for Western music listeners only (Sandell, 1991, p246) although he gave no indication of the geographic background of his subjects, so it may be valid only for the eight NWU students who were test subjects.

When Sandell repeated the experiment with an interval of a minor third between the two tones, he found similar results, with the exception of acoustic dissonance, which was no longer significant (Sandell, 1991, p260). Analysing his results, Sandell concluded that some of the instruments were inherently good or bad blenders (Sandell, 1991, p316).

Considering the effect of onset on blend he wrote "blend will be aided when onsets are either masked, or when the cue that two onsets are present is masked" (Sandell, 1991, p275). This is the principal of common fate again, which Bregman asserts is very important to the perception of blend (Bregman, 1990, pp. 248-292). Bregman also states that blend will increase as part of an auditory stream (Bregman, 1990, p333). This suggests that the type of music played would have an effect on blend. Potentially, music that was chordal could improve the perception of blend, and music that was contrapuntal or solo in nature could decrease the perception of blend. Shepard suggests that the blending of pipe organ notes in an ensemble is partly due to acoustic confusion and partly due to the listener's perception of common fate (Shepard, in Cook (editor), 1999, chapter 10, p117).

2.2.3 Timbral Semantics

Timbral adjectives are most often used when comparing like with like, or for identifying slight differences. They can fall short when people try to use them to describe the differences between distinctly different timbres. Grey (1975, pp. 7-8) identifies the following difficulties with timbral semantics:

- They do not "reveal factors which [are] *uniquely* related to independent properties of the stimuli"
- "A single word may be associated with a number of independent stimulus dimensions"
- They measure "complex aesthetic reactions to stimuli" rather than "direct information about many of the perceptual processes"
- Some words don't exist or aren't available to describe certain perceived differences.

These problems lead some people to dismiss the use of descriptors, as:

> "language is always limited in its ability to express accurately, consistently and unambiguously [timbral qualities]. The results 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" (Creasey, 1998, p51).

However, this inevitability is not demonstrated by Creasey, who cites only two examples (see section 2.2.1). This argument also ignores the fact that descriptors are the primary methods by which we describe sound, even if they are not ideal. Creasey seems pre-disposed towards a vast number of dimensions, defining a timbral space with 335 dimensions, an unspecified number of which are significant. Creasey's timbre space seems to be pre-defined, into which single sounds can be inserted.

This use of single sounds is common to almost every study of timbral semantics thus far. If experiments have only used single sound sources, their conclusions may not hold true for ensembles (e.g. a guitar or piano playing multiple notes, a violin section in an orchestra, a pipe organ with several stops sounding, or a full orchestra). Some adjectives may also only apply to ensembles, or have different understandings in the context of multiple source ensembles.

Sandell defends the use of timbral semantics on cognitive grounds, as they describe a qualitative aspect of timbre. Psychological studies of memory indicate that items are more easily stored and retrieved when they can be put in simple but meaningful categories, such as many of the terms used in table 2.1 below (Rosch, 1975, 1).

Creasey summarises much of the work done to date on timbral semantics from a number of sources (Creasey, 1998, p76). Sandell (1991, p11) also lists the timbral qualities found in orchestral manuals in a variety of categories. The two lists are combined in table 2.1. Research authors are not given here as the author has not verified all the references, but those details can be found in Creasey (1998, p76).

Modalities	Words (from Sandell)	Researched (from Creasey)
Visual	Brilliant, clear, dark, dull, light, pale, transparent, veiled	Bright, clear, dull, penetrating
Tactile	Fuzzy, velvety, smooth, soft, silky, coarse, rough, hard, grainy	Cutting, grating, hard, jarring, rough, soft
Gustatory	Rich, mellow, bland, pungent, tangy	Rich
Geometry, volumes and physical matter	Thick, heavy, thin, rounded, full, hollow, metallic, reedy, brittle, woody, delicate, liquid, glassy	Acute, clang, fat, full, heavy, hollow, nasal, open, reedy, resonant, small, thin
Environmental conditions	Warm, windy, dry, cold, cool	Warm
Moods and emotions	Calm, introspective, expressive, sombre, poignant, melancholy	Intense, laxity, presence

Table 2.1 – A summary of timbral adjectives

Von Bismarck (1974) came up with thirty potential adjective scales (for example fine to coarse, simple to complex and so on). He reduced these to four scales: dull to sharp, compact to scattered, full to empty, and colourless to colourful (in their order of significance). However, the thirty scales were obtained from a variety of different studies, and although they were selected to prevent too much duplication, they weren't obtained from a consistent source, and many of them do seem duplicitous. His scales were limited to steady-state timbres.

Plomp (1976) suggests that von Bismarck's adjectives reveal strong factors but are limited by their inability to describe all nuances of sound. Kendall and Carterette (1993, 1, p452) questioned the cultural significance of Von Bismarck's adjectives, when they found the results difficult to replicate with real instruments. It is worth noting that Von Bismarck's subjects were German, and it is unclear whether the adjectives were given to them in English or German. Kendall and Carterette also pointed out that the use of words in opposition was a potential failing. Dull is not necessarily defined as the opposite of sharp, for example. They developed an alternative system, using precise negatives, such as "sharp" and "not sharp", with greater success.

Kendall and Carterette (1993, 2) then took all the adjectives used in Piston's "Orchestration" (1955) and pared them down until for ten wind instruments, they were reduced to two scales. These were "nasality" versus "richness" and "reediness" versus "brilliance" for that timbre space.

Within the musical acoustics community, certain words such as "harsh" have a precisely defined meaning (in this case, referring to sounds within a critical bandwidth – see p83 in Cook 1999). Bregman suggests that of the various timbral descriptors, brightness is perhaps the most distinct dimension of timbre (Bregman, 1990, p646). Brightness is commonly understood to relate to the frequency of the spectral centroid.

One potential problem of timbral semantics is that many words simultaneously describe both loudness and timbral quality (Sandell, 1991, p25). This suggests that perceived amplitude is an important factor in choosing a particular word from a library of possible semantics. This matches the way in which timbre change is an important perceptual clue to the volume of an instrument. Close instruments being played softly are easily distinguished from instruments being played louder but more distantly, excluding other considerations (Chowning, 1999, pp. 269-274).

Players may use words to describe both the sound of an instrument and the feeling of playing it, but their understanding may not be consistent for both meanings. For example in Geissler et al. (2003) a large number of German violinists identified various characteristics they would use to evaluate a violin. Some words used were timbral (bright versus dark, warm versus cold, round versus soft) – these made up the largest portion of replies at 38%. Others, such as easy versus difficult, came under the heading of "reaction", perhaps better contextually translated as "interaction". It is interesting to note that this categorisation is applied retrospectively to the results by the researchers, as indeed have the words in table 2.1 by the author. It may be interesting to see how the listeners who generated the words in experiments such as Geissler et al. would rate them. Other researcher-generated categories with over 10% of words in were "sustain" and "balance". Geissler went on to use seven of the words, selected for their breadth of coverage, rather than those that were most commonly used. These words were "bright", "passionate", "nasal", "pleasant", "reaction", "Balanced", "colourful". "Pleasant", "reaction", "bright" and "balanced" had common understanding among the ten subjects used. "Passionate", "colourful" and "nasal" did not have common understanding.

Several other similar experiments have been done. One problem is that a certain set of adjectives will only work for some subject groups, and it is possible that some descriptors will have different understandings in different contexts. A group of organists may have come up with an entirely different set of words. It is also clear that there are some difficulties in translation, as the words generated by von Bismarck and Geissler et al. do not all seem obvious when translated. This may be because the words have been poorly translated, or because a literal translation does not have the same nuances of interpretation across language barriers. What is clear is that while research in different countries can provide useful indications of common understanding, it cannot always provide precise definitions outside of the language or subject group used in the tests. Because of this, attempts to translate terms like for like are likely not to be of great use. Creasey (1998, p86) suggests that the range of perceptible timbral nuance far exceeds linguistic ability even within a single subject and language group.

With those caveats in mind, several experiments across various language groups are worth noting at this point. Moravec and Stepanek (2003) asked 109 Czech musicians for adjectives they used when describing timbre. 1964 words were collected remotely, with no stimuli being presented. Each respondent had at least one word unique to himself or herself. The commonest 30 words were (in descending order): sharp, gloomy, soft, clear, velvety, round, delicate, unpointed, hard, bright, harsh, sweet, full, dark, rough, warm-hot, radiant, warm-hearty, clear, coloured, ringing, lucid, narrow, wide, cool, metallic, cold, shining, blurred, smooth.

The frequency of particular word occurrence varied depending on the instrument the respondent played. This supports the previous suggestion that different words are important in different contexts. While the language difference is important to note, the researchers were keen to emphasise similarities to other studies with native English speakers. No single word was used by every respondent.

Nykänen and Johansson (2003) gathered Swedish words from saxophone players by both asking for words, and also playing examples (recorded and in person) of different saxophones. The ten most common words are here translated to nearest English equivalent, with alternative English suggestions after discussions between the author and Arne Nykänen. They are: Large (big or fat), rounded, rough, warm, soft, nasal, coreful, sharp, and "having bottom". The latter is probably best translated as "being well supported" or "strong". The word "coreful" is intriguing as in its original form it doesn't exist in Swedish, but the subjects used it nonetheless. The best suggestion for an English translation thus far is "having body or substance".

The results were analysed in a PCA, which took as its possible components the fundamental frequency, formant characteristic and frequencies of the first 6 formants, spectral width, the four well defined descriptions of loudness, roughness, sharpness and tonality, the type of saxophone, and various listener variables. Correlation was found between the first two dimensions and sharpness and roughness respectively.

Leman et al (2003) rated a number of different musical excerpts, using 100 listeners in Belgium. It is unusual in this context as short extracts of music by various ensembles were rated rather than individual sounds. The researchers chose fifteen contrasting scales as a rating method, such as Tender to Bold, or Hopeful to Desperate. These were then put in three groups – Valence (favourable vs. Unfavourable), Dominance (relating to power and control), and Arousal (i.e. exciting vs. boring). This again shows translation problems, as several of the direct translations were literal and in English had subtle, or in some cases, unsubtle and unfortunate alternative meanings. Valence was related to the number of onsets and inter-onset interval, dominance to roughness in most cases, and arousal showed no clear results. There does appear to be some common understanding of a relationship between the words used and the music. While this involved the use of ensembles, people here were not specifically describing the sound, and their responses were clearly more emotive than analytical from the titles of the rating scales.

No-one has yet examined the way in which the common understanding apparently present in some papers has been acquired by the subject group. Levetin, (in Cook, editor, 1999) notes that semantic memory remembers the definition of words, but not when or how that learning occurred (p211). Yet if particular sociological groups have common understandings, and as has been shown, that understanding has at least some variation from one language to another, there must be some learning process. Outside of a formal environment such as an organ builder's workshop, it is hard to determine where that process takes place, as consideration of timbral semantics is not directly taught as any part of a musical education.

2.2.4 Other Perceptual Effects

There are a variety of other perceptual effects it is interesting to note at this point. Timbre can be affected by phase (Bergeijk et al., 1960), and this would particularly be the case in pipe organs, where combinations of stops have the potential for phase-related cancellation. Inharmonicity also has an effect on timbre (Fletcher et al., 1962). Intriguingly, Charboneau (1981) was able to successfully simplify (in data terms) the Grey reference tones by reducing the inharmonicity. He concludes that inharmonicity or quasi-periodicity existing in natural sounds is not intrinsically significant to the timbre of instruments playing continuous tones. He qualifies this with the proviso that this probably isn't true for percussive sounds. The ear is much more sensitive to the frequency variations of all the harmonics than to whether or not these variations are different for each harmonic.

He also found that simplification of onset and offsets worked well. If particular individual harmonics were late in onset or early in offset, normalising them to the onset and offset of other harmonics didn't make much difference to their perceived timbre. Auditory streaming has been referred to in section 2.2.2. Scene Analysis (Bregman, 1990, chapter five) divides a complex auditory environment into melodies and other events linked in streams of some fashion. This streaming can be by pitch, timbre, melodic pattern and/or rhythmic pattern. Gregory (1994) verified that timbre was an important aid in separating musical streams, and by using MDS techniques came up with a selection of timbre dimensions applicable to his sample set.

An effect related to streaming is grouping, which considers how timbres, rather than auditory streams with a common timbre, are classified in the brain. Such grouping is not by artificial classification, but by features of tone. If a listener hears a flute, their immediate reaction is not to think "woodwind" but to think "flute tone". This brings into play the concept of timbral hierarchies, or how the brain classifies or attempts to structure timbres. In this theory, prototypes (e.g. red) attract other non-prototypes (e.g. pink or orange) to themselves. This means that pink is perceived to be closer to red than red is to pink (Rosch, 1975, 2). Such classification is difficult to formalise but is potentially of use when understanding subject responses.

Vibrato is a tonal modifier more often found in singers than in pipe organs. It can be introduced in pipe organs with a device known as a Tremulant. In any context, a large vibrato is perceptually disturbing. However, for some instruments, such as the violin or the human voice, a lack of vibrato encourages fission (the hearing out of spectral components) rather than fusion, and this lack of vibrato can have a grating effect (Lerdahl, p142, 1987). This does not appear to be the case in the pipe organ, perhaps because listeners are more familiar with it in its untremulated state. Tremulants are rarely used in general pipe organ music, although they are called for in certain specific circumstances.

Pollard and Jansson (1982) developed the tristimulus analysis mentioned in section 2.2.3. This method represents a single spectrum as the ratio between the fundamental, the sum of harmonics 2-4, and the sum of the remaining higher harmonics. The perceptual theory behind this technique is Pollard and Jansson's assertion that those regions represent the broad categories of information to which listeners attend to make distinctions between spectra. When Sandell used this technique to compare spectra, and used regression to analyse the result, he got this equation:

$$\text{blend}=-0.067\text{fundamental} + 0.111\text{mid} + 0.415\text{high} + 0.306.$$

This may seem crude, and Sandell did not suggest that it was a comprehensive description of blend, but it goes some way towards indicating the relative contributions of each frequency band to overall blend. Perceived blend again seems particularly related to the higher harmonics rather than the lower harmonics, supporting the theories developed in section 2.2.2.

2.3 Simulation of the pipe organ

The pipe organ is one of the most widely simulated classical instruments, perhaps second only to the piano. This may be because it is easier to replicate than other instruments, or because it is the most expensive and unwieldy instrument in its natural form. Discussion of simulation, in particular the Bradford system, is included here because it is proposed to use a simulator as a useful method of verifying theories later in this project.

Assuming that theories of relationships between timbral semantics and acoustic phenomena are developed, they will have to be verified by the creation of examples that should exhibit those characteristics. Using a pipe organ to achieve this would be expensive and impractical, and might risk permanent damage to the organ. The Department of Electronics owns a Wyvern B235 electronic organ with voicing software, based upon the Bradford technology introduced in section 2.3.1, which is ideal for the proposed usage.

There are two technologies used for simulating pipe organs in current commercial use. The first uses sampling technology. Recordings are taken of the stops and stored in a wavetable. The advantage of this method is that individual stops can be imitated very well. The disadvantage is that the sympathetic effects of combining several stops cannot be imitated, and the samples are simply summed together with no interaction.

In a pipe organ, stops drawn together interact in various ways in addition to the readily apparent summation of their individual outputs. Two stops very slightly out of tune may "draw" together such that they speak at the same pitch. This can happen with pipes speaking at harmonic multiples, and helps to draw the pipes into a coherent ensemble. Pipes may also interact through sympathetic vibration of the windchest and air in the wind-channel. While this interaction is not always desired in real life, its simulation where the effect is desired is an important part of creating the illusion of reality.

The second method currently in commercial use involves synthesis of ensembles, and is that used in Bradford based systems. The simulation of individual stops is more complex, as they must be synthesised from analyses of real stops. However, the synthesis method of tone generation makes it easy to take account of the various effects of real organs as stops are combined, so these organs typically excel in their ensemble. Changes to a stop's harmonic composition can be easily achieved, unlike sampling systems where filtering provides a more limited palette of adjustment.

With any simulation, the goal is to get away from the electronic origin of the sound, and convince the listener that they are hearing the real thing. Reverberation, particularly naturally occurring, can help in this illusion, as "the impression of spatial distribution is increased by the reflections to which the simulated sound is subject in [a] reverberant building" (Kitching and Comerford, 1988, p124). An equally important factor is adequate sound dispersion, so that the electronic instrument can begin to approach the inherently multiple source nature of the pipe organ.

2.3.1 The Bradford Musical Instrument Synthesiser

An overview of the BMIS, with specific regard to pipe organ simulation, is given in Comerford (1993). A basic technical introduction can be found in Comerford (1981). The generators (devices that produce the notes) are 24-bit pitch counters. There are 64 on each board in the current system. Each generator is allocated "on the fly" so each note when played is allocated a generator per stop in use at that time. Eventually there would be no generators left, so some sharing of generators is necessary. Stops that share a particular generator are phase-locked together to avoid cancellation effects, but otherwise there is no phase locking – this helps to combat the problem of electronic instruments always sounding the same. There are algorithms for the selective dropping out of notes based on auditory masking when the system runs out of generators. For example, a Dulciana might share a generator with a Diapason, but this must be done carefully to avoid losing the perception of an ensemble. (Comerford, 1987)

There is also a controlled degree of random variation introduced in parameters such as pitch and phase. The BMIS can also simulate inharmonic components, if necessary assigning a generator to each partial (Kitching and Comerford, 1988, p121). While the BMIS is optimised for recreating pipe organ sounds, it can simulate others, although sounds such as a piano take seven generators to accurately simulate each single note.

BMIS has both multiple cycle and single cycle modes. In single cycle mode, each time a stop is changed, the on-board processors calculate the new waveform at each of the voicing points on the keyboard. Keys in between the voicing points have their waveforms calculated by interpolation. In multiple cycle mode, more cycles of the waveform are stored in the onboard memory, including complicated parts like the transients. The parameters for the multiple cycle waveforms are taken from samples of real organs, so to some extent it acts as the best possible combination of synthesis and sampling organs.

Multiple cycle technology only needs one generator per note, so is relatively efficient at using generators. This is especially useful in the transients, where complex waveform transients no longer have to have a high generator overhead. Some stops in the single cycle mode use two generators, one for low frequencies (typically the first two harmonics) and the other for the higher frequencies, as well as more generators for the transient. One disadvantage of multiple cycle technology is that more memory is required on the system boards than for single cycle voicing, so there is a limit to the number of multiple cycle voices that can be implemented at one time.

2.4 Conclusion

In this chapter, research and literature relevant to the proposals introduced in chapter one has been overviewed to provide a firm base for the remainder of this thesis. While mostly concentrating on the pipe organ and timbral semantics, other research of possible relevance has also been covered. The limitations of existing work and areas in which little has been done have been noted, giving potential for future work both within and outside the scope of this thesis.

3. Introductory Experiment

In this chapter, the initial research strategy suggested in section 1.5 is implemented and developed. For an initial experiment conducted as part of the proposal for this thesis, the theories developed in Disley (1999 and 2000) are used to test the suggested method of verification by synthesis. The results of this experiment are used to reconsider the experimental procedure and analysis methods suggested in chapter one.

3.1 Overview of theory behind the initial experiments

As a result of analysing acoustic ensembles, Disley (1999 and 2000) proposed the theories about perceived "blend" and "strength" described in section 2.2. In summary, perceived "blend" in a principal ensemble was related to a smooth tail-off of harmonics from low to high with no prominent peaks or troughs. Perceived "strength" was related to a solid unison (or sub-unison) foundation and a rich harmonic development reinforcing the harmonics of the unison.

When theories of timbral semantics are developed, it is important to prove that theoretical qualities or descriptions can be matched with actual qualities that can be synthesised. While this is not an ultimate verification of the theory, it is another step in that direction and one appropriate in cases such as this. Disley (2000) verified and developed the theories using carefully selected recorded examples with characteristics expected to elicit a certain subject response. In order to verify these theories further, synthesis theoretically offers the opportunity of creating samples that vary only in desired ways and in no others, depending on the synthesis technology used. As in Grey (1977, p1271) synthesised samples are much easier to manipulate and control precisely. It is simple to create an entire experimental sample set with the same tuning, the same temperament, and the same reverberant environment. This is something that is difficult to do when using recordings of real pipe organs.

The question to be answered at this stage is: within the restrictions of a controlled experiment, how realistic or how artificial is tolerable? Efforts must be taken to minimise any sensation of artificiality. Churcher (1970) suggests that electronic tone is less offensive to the ear when no direct sound is heard, and when it is in a reverberant environment. It is easy to add controlled identical reverberation to all of the samples. Even with this, any synthesised example is likely to be identified as electronic. In previous experiments (Disley, 1999, ch.7 p18) subjects not only identified an electronic organ as such, but even thought that recordings of real organs sounded artificial. Fletcher et al. (1963) discovered that musicians were much more likely (than non-musicians) to falsely identify real recordings as electronic.

A large number of experiments have been conducted where auditory stimuli have been constrained to vary only in known ways (for example, Howard and Silverman, 1976). The problem with such experiments is that the results are often unmusical. This is because to fully engage the auditory system, sounds should be complex in frequency (for example having spectral variation during the onset transient), and complex in time, with regard to both amplitude and spectral envelopes (Grey, 1975, pp. 1-15 and Handel, 1989 pp. 226-263). Or, as Hall (1993) describes it:

> "Truly musical stimuli are complex successions of sounds, heard in continually changing context"

In this case the aim is to explore theories related to the timbre space of the pipe organ. The synthesised samples must bear close resemblance to the sound of the pipe organ, otherwise this experiment would not be dealing with the timbre space of the pipe organ but with a different and artificial timbre space. When considering whether to use real or artificially generated samples, it is important to bear in mind that the results will be limited by constraints imposed at this stage. If synthesised samples are overused, an experiment will not prove that the qualities measured are necessarily primary perceptual attributes of listening to pipe organs.

In the case of the previous experiments that developed the theories on blend and strength, the words "blend" and "strength" were not given any precise definitions. The problems associated with providing a definition have been discussed in section 1.5.3. As the effect giving a definition of the terms being used as rating scales is unclear, the opportunity was taken to use the preliminary experiment to divide listeners into two groups, one of which was given a definition of "blend" and the other one of which was not. The results might indicate whether such a definition would reduce experimental noise, or if it would just guide subjects towards a particular conclusion.

3.2 Synthesis of examples

To verify the theories on blend summarised in section 3.1, four different samples with varying levels of theoretical blend were synthesised to be used in a comparison test. For use in verifying theories developed with real-world examples, the control benefits here outweigh the removal of the tests from a real-world timbre space. As these samples would be compared with each other rather than recordings of real pipe organs, it wasn't necessary to base them on any particular pipe organ, although for the reasons elucidated above the samples did have to remain within the timbre space of the pipe organ. The four samples were instead based upon the default specification of the Bradford synthesis based B235 instrument described in section 2.3.1 as supplied by Wyvern. This default specification uses the standard single-cycle voicing technology rather than the more complex multiple-cycle voicing proposed for verification of theories developed in this thesis.

As the theories had been developed in the context of full principal ensembles, the synthesised examples consisted of the four principal stops: Open Diapason 8', Principal 4', Twelfth 2 2/3' and Fifteenth 2'. The mixture stop was not included as discussion with staff at both Wyvern and Bradford University provided some confusion over how mixtures were simulated, particularly with regard to implementation of break points. Revoicing of the synthesised examples was done with the aid of Wyvern's Digital Enhanced Voicing (DEV) package. This uses a custom PC card to interface with proprietary ports on the circuit boards of the instrument, not usually available to customers of the commercial product. DEV is a package developed in-house by Wyvern, and is not as robust or comprehensive as a commercial product would need to be. It is also comparatively poorly documented, so much of the information about its use in this section is derived from the author's use of the program and communication with the program's author at Wyvern. DEV permits control of the steady state harmonic spectrum for each stop at a number of voicing points across the range of the keyboard. Notes that are not voicing points derive their parameters by linear interpolation between the voicing points either side of them. It is also possible to balance the overall levels of each stop in the full specification of the organ.

What DEV cannot do is to alter the transient portion of the sound or the precise harmonics present at a given voicing point for a certain stop. The latter can only be altered by manually editing the associated proprietary format specification files. The transients can be edited in a separate programme called Envelope Studio, which offers an alternative and more complex interface to the specification files with its own limitations. Here only one stop and voicing point can be viewed at one time, but each harmonic can have its amplitude varied over the period of the onset transient. Some parameters remain unalterable in both programs.

Revoicing of all the samples for this initial experiment was done in DEV, as its ability to display multiple windows made comparative editing easier. The first of the four ensembles was simply the default specification of the organ, with no editing. The three variants were developed in the following ways.

The author did the first revoicing by ear to tailor the chorus more to his preference of a good ensemble. This ensemble is referred to as the "revoiced" ensemble, and mainly involved adjusting the overall balance of each stop to create a more cohesive chorus. The second revoicing was designed to significantly reduce the upper harmonics of all component stops. Based on the theories outlined in sections 3.1 and 2.2, this should increase the perception of blend. This ensemble is referred to as the "dull" ensemble. The third revoicing took the second, "dull" specification and increased the amplitudes of the harmonics that were exact octaves above the fundamental (harmonics 2, 4, 8, and 16). This increased the frequency of the spectral centroid in a subtle manner suggestive of adding more upperwork. A smooth roll-off of harmonics is a requirement in the theories being tested. Increasing these isolated harmonics would test if the smooth roll-off was indeed necessary for a well-blended ensemble. This final revoiced ensemble is referred to as the "bright" ensemble.

As altering the transient spectrum within the voicing software was difficult and limited, and the four examples were all similar, the transients were not altered apart from adjusting the final volumes to match that of the steady state waveform. Thus the major difference between the four samples was in the steady-state part of the sound. It should be noted that the alterations to each stop were made in their context as part of the overall ensemble, and without concern for their solo use. This was a valid methodology as these ensembles were only used with all four stops together.

Ideally, the four ensembles used in this test would be presented using the Acoustic Signature graph method later used in analysis of real ensembles and described in chapter five. This relies on Fourier analysis of note-by-note ensemble recordings over the keyboard range. Unfortunately, when the attempt was made to do this, the experimental organ began to malfunction and continued to do so thereafter. These problems are described in more detail in section 6.1.1. However, the voicing software saves its data as harmonic amplitudes, and thus a theoretical composite is possible by extracting the data and summing each stop's contribution to each harmonic. This cannot be directly compared with the Acoustic Signature, as it takes no account of the relative volume of each stop or the effect of the synthesis and amplification equipment on the sound. While the relative volumes of each stop are included in the overall files for each revoiced specification, it proved impossible to extract these in the way that harmonic data for each stop was extracted. Only the "revoiced" ensemble had its relative levels altered, and this has not been included in the graphs below as its initial harmonic levels were identical to the base specification.

Comparative graphs of the three remaining synthesised ensembles can be found in figures 3.1, 3.2, 3.3 and 3.4. Each shows the harmonic data for one of the four voicing points corresponding to the range of the musical example used. In the keys beside these graphs, "b" refers to the base specification, "d" the dull specification, and "f" the bright specification. Numerical data for each specification can be found in appendix A.

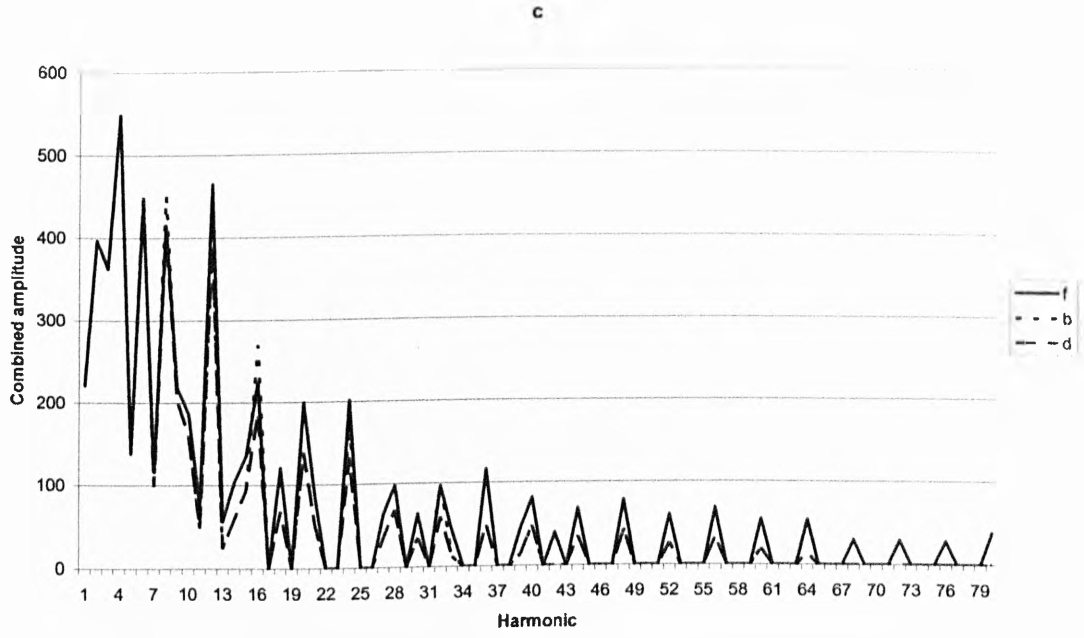


Figure 3.1 – Comparative harmonic data for voicing point c

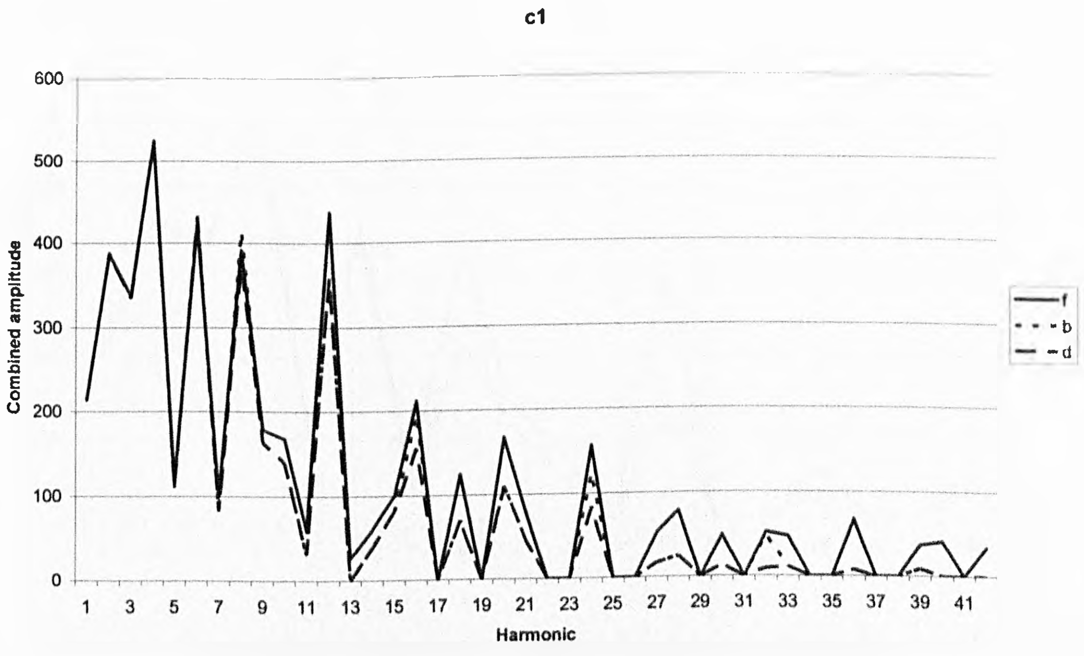


Figure 3.2 – Comparative harmonic data for voicing point c1

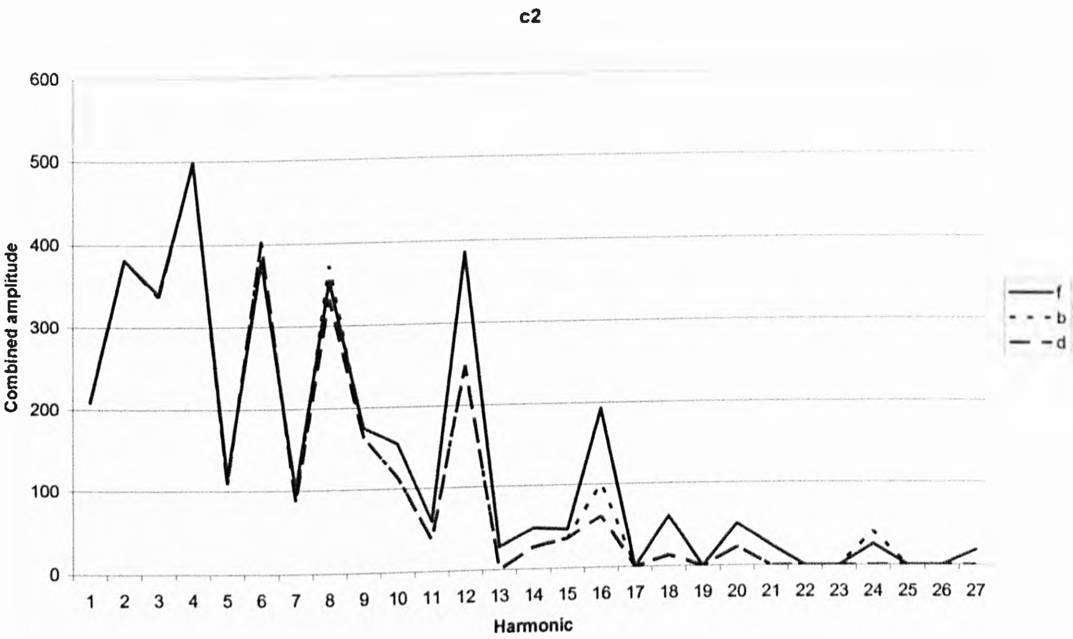


Figure 3.3 – Comparative harmonic data for voicing point c2

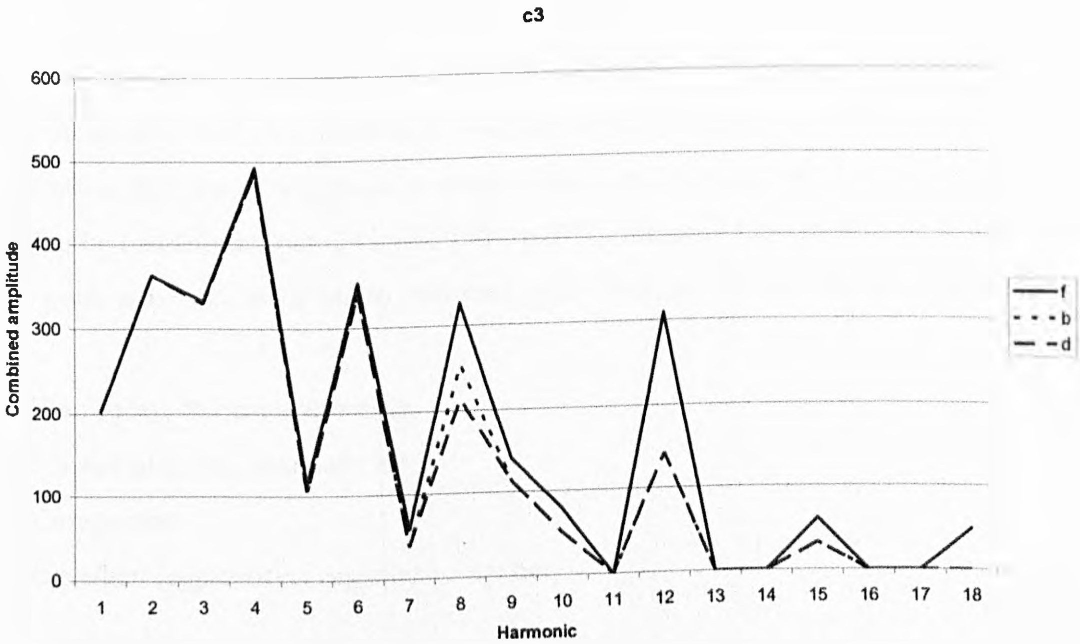


Figure 3.4 – Comparative harmonic data for voicing point c3

3.3 Selection of Subjects

Selection of subjects for an experiment where specialist knowledge is needed is never easy. The specialist knowledge in this case was familiarity with the timbre space of the pipe organ, and sufficient musical ability or listening experience to be able to identify and comment on the small differences between the four synthesised examples. As such a subject group is rare, using the techniques described in section 1.5.4 the Internet can be used as an experimental delivery method to increase the catchment area.

Some authors have argued for the exclusive use of musical test subjects. Miller and Carterette (1975) suggest that musical subjects are important in timbral experiments as they have a more stable space of perceptual dimensions. However Eagleson and Eagleson (1947) states that professional musicians are worse than more generally musical people at discerning timbres, perhaps because they are too specialised at listening critically to their instruments. In this initial experiment, instead of selecting only subjects with a certain level of musicality, all subjects will be invited to self-classify their musicality. Self-classification is necessary as where subjects are not known, and even where they are, it is difficult to assess a standard of musicality, and impracticably lengthy by formal methods (Mursell, 1964, p287). To make this classification simple, a four point scale was used, where subjects could select one of the following options:

1. Don't play, but enjoy listening
2. Some ability (e.g. manuals only)
3. Competent
4. Excellent (e.g. concert organist or teacher)

Self-selection does however have the potential for errors. When this test took place, the author saw two different people of wildly differing standards choose the same "competent" rating, one of whom he would have given a lower rating, the other a higher. However, the final self-classification (in table 3.3) included entries in all categories, and the significance of subjects' musicality will be considered in section 3.7.

A second consideration when selecting listeners is how many are required for the experiment to have significant results. Too few for a particular scenario will result in the end product being biased. While ideally the number of subjects would be as large as possible, the difficulty of getting subjects to take part means that testers aim for a realistic minimum of listeners, particularly where a requirement of specialised knowledge reduces the number of potential subjects. Levetin (in Cook, 1999) suggests a minimum of between five and ten listeners for phenomena that are expected to be relatively invariant among listeners, and between 30 and 100 for phenomena where large individual differences are expected.

It is not entirely clear which category these experiments fall into, as it is hoped to find common understanding but this is not certain. For the initial experiment, fifteen subjects were gathered, somewhere between the numbers suggested above, but not inappropriate for a pilot study. Potential subjects were gathered both from an Internet pipe organ mailing list (piporg-l), and from people whom the author knew to have an interest in pipe organs. Subjects were also surveyed as to their age band and their geographic location. The results are given in tables 3.1, 3.2, and 3.3 below.

Location	Subjects
Canada	1
Holland	1
United Kingdom	4
USA	9

Table 3.1 – Geographic locations of test subjects for the initial experiment

Age groupings	Subjects
Under 25	2
26-50	5
51-75	7
Over 75	1

Table 3.2 – Age range of test subjects for the initial experiment

Musicality	Subjects
Don't play, but enjoy listening	1
Some ability (e.g. manuals only)	3
Competent	7
Excellent (e.g. teacher/concert)	4

Table 3.3 – Self-selected musicality ratings of test subjects for the initial experiment

The relatively small number of participants in this initial test was due not to a difficulty in finding subjects, but a bug (or undocumented feature) in later releases of the commonest Internet web browsing software. Previous Internet tests by the author had used a "mailto" form submission feature, but support for this feature was quietly removed from later software releases. Thus many volunteers attempting to take the test who had upgraded their web browsing software found themselves unable to submit their answers. Details of the "mailto" feature, and the alternative method used by the author in subsequent tests, can be found in section 3.4.1 below. Fifteen listeners was considered adequate for an initial experiment, so extra volunteers were not recruited until the next stage of this study.

3.4 Experimental procedure

Comparison tests have been introduced in section 1.5.4 as the only practical method if remote testing is to be used. The question then arises of what scales to use. Clearly if a word is being tested, it should be the label for a comparative scale, but what resolution should that scale be? Should a user be presented with a continuous scale or one with a number of discrete points? The extreme example would be a three-point scale, where the amount to which one example exhibited more of a quality was not indicated. Rating scales could also be presented to the used in more than one dimension.

Presenting a listener with a rating scale of too small a resolution would result in discrimination being lost. Presenting a listener with too great a resolution of scale when they are attempting to quantify imprecise percepts would result in unnecessary confusion to the listener as they attempt to pick a precise answer, and increased noise (random variation) in the analysis of answers. Use of a continuous scale simply delays the decision of resolution, as answers must be quantised later to a desired number of points.

The use of multi-dimensional rating scales brings in a new set of complexities. Each additional dimension creates a new difficulty in giving the listener a method of altering those dimensions and then representing the values on them to the listener. A practical limit would be four dimensions, represented by, say, a three dimensional visual display with colour or texture as the fourth dimension. A less involved experiment would be practically limited to two variable dimensions, presented for example as an x-y graph on a computer screen or piece of paper.

This necessity to limit the number of dimensions causes problems when investigating timbre, as it implies that the titles given to the dimensions used represented the limit of timbre space. This suggestion of a low and finite number of timbral dimensions is not necessarily the case, as in the arguments of Creasey (1998) summarised in section 2.2.1. Such a reduction in scope of timbral dimensions (so-called cartoonification) can be useful in understanding or demonstrating basic concepts, but is less useful when precise understanding is required.

A multi-dimensional approach to testing requires a different test set-up from comparative testing. With a single scale comparative test, two samples can visually be placed at either end of the scale, and the scale used to indicate which sample has more of a particular quality. With a multidimensional approach, the user must place samples in the timbre space, and the scales must therefore vary from "possesses a minimum amount of this quality" to "possesses a maximum amount of this quality". Both methods have the potential for different perceived scales from listener to listener. Some may use the maximum amount of scale given, whereas others may concentrate their answers within a small part of the scale. This allows room for additions to the sample set that may exhibit greater levels of a particular quality than those already rated. Whatever the test methodology chosen, it is wise to examine how test subjects are responding to and using the scales available to them.

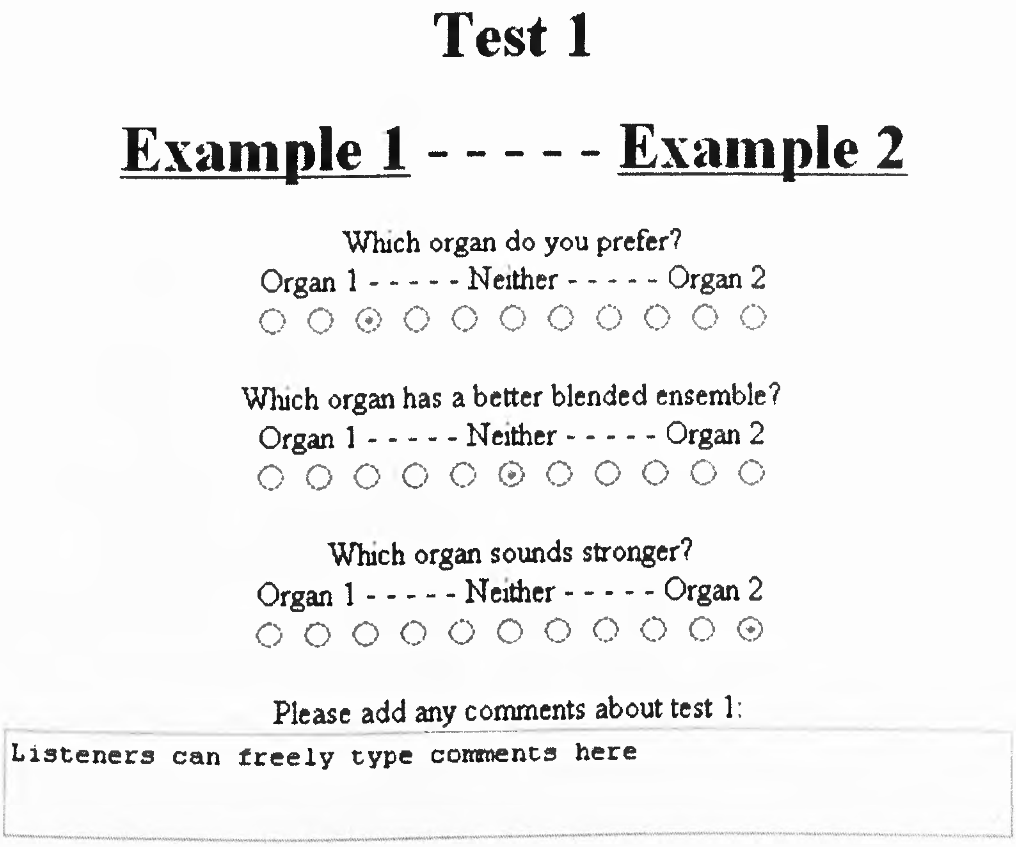
3.4.1 Implementation of psychoacoustic tests over the Internet

The use of the Internet as a testing medium also brings into consideration the method by which the test is administered. An Internet web browser renders HTML, which has a limited scope as a test interface. It is possible to implement more complex interfaces through browser plug-ins, such as the programming language Java or the proprietary multimedia Flash environment. Both these and other similar browser extensions add to the complexity of test creation, and bring with them questions of compatibility.

Within HTML is a simple method of gathering information known as a form. A form can consist of lists, from which it is possible to select one or more options, check boxes, radio buttons (like check boxes, but which permit only one of a set to be selected), and text input areas. There is no option for a continuous control. Submission of the form is by the user pressing a button on screen that then implements one of a number of commands. The simplest of these is the "mailto" feature, which emails the contents of the form to a given email address. The author in his previous work used this combination of HTML form and mailto, but successive generations of web browsers have poorly implemented the mailto feature, usually now resulting in a blank email message being presented to the user and no data being submitted.

Alternative methods of form submission involve the execution of code (in PERL or similar computer languages) held in a CGI directory of the web server. While needing some programming skill, the simple processing of text-based input is not complex. At the University of York, the Department of Psychology has set up a simple Web Response system that processes, stores, and tabulates input from web forms. Through custom code inserted in the web form, it can force users to complete all parts of the form before submission. It does not offer any automatic processing of results, but can redirect the test subject to another web page to confirm that the data has been successfully submitted, another feature lacking from the mailto form submission method. Permission to use this Web Response system was granted to the author by the Department of Psychology's system administrator, Rob Stone.

Use of this form of testing interface is simple and practical for both the person taking the test and the person creating the test, provided that the test can readily be adapted to the simple web form interface. This precludes multi-dimensional testing and leaves single dimension comparative testing with a quantized scale as the best form of web testing to use. As few people have thus far explored this way of testing, there is little literature to recommend how many points such a scale would have. Eleven distinct points were used in the author's previous research, to give a variety of points including a "neither" or "both the same" option in the middle, but without an overly complex number of points. As this worked well in those experiments, it will be used here, the scale again implemented as a set of eleven radio buttons, which automatically prevent selection of more than one option on the same scale.



Test 1

Example 1 - - - - - Example 2

Which organ do you prefer?
Organ 1 - - - - - Neither - - - - - Organ 2
○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○

Which organ has a better blended ensemble?
Organ 1 - - - - - Neither - - - - - Organ 2
○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○

Which organ sounds stronger?
Organ 1 - - - - - Neither - - - - - Organ 2
○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ●

Please add any comments about test 1:
Listeners can freely type comments here

Figure 3.5 – Web form as presented to listeners

The length and resolution of samples presented to the listeners in the tests was another matter of debate. From a conventional listener's perspective, the samples need to be as long and high resolution as possible. The length gives the listener time to appreciate the finer nuances of the sound, and the resolution ensures that the listener is hearing the full bandwidth of the sound. However, testing over the Internet puts these two requirements in conflict with the time taken to download large sound files. A large sound file will not only discourage listeners from taking part, but increases the risk of unsuccessful or incomplete file transfer to the listener's computer prior to attempted playback. In some cases, listener's computer systems prevented them from downloading a file of greater than 1MB, and even for those without that limit, 1MB per sound file is a sensible limit to avoid overly lengthy download times (1MB will take about three minutes to download over a currently typical 45kB connection).

Madison and Merker (2003) suggest that listener ratings for more general terms, such as "swingy" or "simple", reach a fairly consistent level after playback of slightly less than 2 seconds in most cases. This was using very abstract terms describing the overall style of the piece of music rather than the precise sound they were hearing. If the entirety of an organ's ensemble is to be tested, the musical example must be long enough to demonstrate a number of different notes. The example must also be in a musically appropriate style so that the instrument is being demonstrated in the manner it was designed for. The author's previous experience also suggested that listeners are better able to appreciate an organ in its environment when the reverberant portion of the sound after the last note has been released is also presented to them.

One way to reduce the size of the sound file is to use some form of compression. However, there are no loss-free compression algorithms in common use. Some early compression formats rely on reducing the resolution of the file to eight or four bits instead of the 16-bit sound of CD quality. This introduces increasing levels of noise as well as decreasing the accuracy with which the audio waveform is reproduced. It is possible to achieve reductions in file size by use of the loss-free compression method used in ZIP files (and GIF/PNG images). With text or image files, large reductions in data can often be made, but as this reduction is only about 20% in the case of sound files, no such equivalent sound file format is in common usage. The only way to achieve this reduction would be to supply the remote listeners with a compressed file they had to decompress themselves. This is obviously not desirable, and would prevent the test from being easily taken via a web page.

More recently, computationally intensive psychoacoustic compression algorithms have been developed. The most common of these is the MP3 format, which is similar in principal to Microsoft's proprietary WMA format and Ogg's Vorbis format, as well as the method used in Minidisc recording. All of these are lossy compression methods which provide large reductions in file size by eliminating parts of the data stream that are less important to human perception of the sound. They work particularly well for popular music, but at the more extreme compression ratios a change in the quality of sound is particularly noticeable. They are the auditory equivalent of JPEG compression in the image world. It may be the case that some of the higher resolutions of this type of compressed file are appropriate for psychoacoustic testing. However, until research has been conducted into what the realistic minimum bit-rate is for lossily compressed files to be used validly in psychoacoustic testing, such use is to be avoided.

If lossy compression cannot be used, and files must maintain their amplitude resolution (typically at least 16 bit) and to ensure the listener is presented with all relevant data, other forms of compression must be considered. File length is the obvious candidate, so the file should be no longer than necessary to present an appropriate musical example.

One way in which the amount of data can be halved is to present examples in mono. Stereo cues are not necessarily important in timbral experiments (Goad and Keefe, 1992, p45). Mono sound reproduction is particularly appropriate for the pilot experiment, as the artificially generated samples from the Bradford synthesis system are split between the left and right channels, alternating each semitone. This mimics one particular layout of real pipes, and works well with external speakers, but is unrealistic over headphones. This simple system of stereo, with its relatively basic integral artificial reverberation unit, may also benefit from being presented in mono, as listeners may be less able to discern its artificially generated nature.

In the author's previous research, however, it appeared that presenting examples of real organs in mono had a significant effect on listeners' perceptions of the sound. Some degree of cancellation could not be avoided, so this process potentially altered the sound. Certainly the listener's perceived image of the organ in its acoustic environment is altered by reduction from stereo to mono. This leaves only alteration of the sampling rate as the final means of reducing file size. Ideally, the sampling rate should be high enough that the entire range of human hearing is covered, and the CD sampling rate of 44.1kHz is the standard for such a resolution. In practice however, the upper limit of human hearing deteriorates with age from the theoretical limit of 20kHz (Hall, 1980, p105).

Additionally, acoustic analysis of the samples recorded from real pipe organs used in later stages of this thesis showed diminished acoustic energy at higher frequencies, with little or none in the uppermost octave. This is not the case in much popular music, but appeared consistent over several pipe organs. This may be due to the fact that high frequencies are the most readily absorbed, and thus in the distance recording setup required to obtain the typical sound of a pipe organ, the high frequencies never reach the microphone. Provided the microphone is placed in a typical listener position, this lack of very high frequencies is not a problem as the listeners would not be able to hear them either.

Hence reduction of the sampling frequency from 44.1kHz is valid for recordings of real pipe organs. The next widely supported sampling rate is 22.05kHz, which results in a practical upper frequency cutoff of approximately 10kHz. Where stereo examples were used, this down-sampling to 22.05kHz provided sufficient reduction in file sizes (50% compared to an otherwise identical sound file recorded at 44.1kHz) to keep the files below the 1MB limit described above.

3.5 Test Procedure

When setting up an experiment, the question should be asked: "What is an appropriate task for subjects to undertake?" For this preliminary experiment, the subjects were asked to compare several pairs of musical examples using labelled scales to indicate which, if either, of the two samples had more of a certain timbral quality.

The scales used were "blend" and "strength", the two timbral descriptors for which theories had been developed in the author's previous work, as well as "preference". This last scale gave the listeners the opportunity to indicate a personal preference for one or the other, and was included more for interest than because a particular correlation between preference and the other ratings was expected. Personal preference, positive or negative, could be associated with the use of certain timbral semantics. Opportunity was also given for free comment by means of a text box.

As it was not clear at this stage whether a formal definition of the terms used as labels for ratings scales was desirable, half of the subjects were given this formal definition of blend before they commenced the test:

> "the tendency for concurrently-sounding timbres (for example the many pipes in a pipe organ) to fuse into a single timbre"

Those who were given this definition were asked whether they agreed with it. Those who were not given a definition were asked at the end of the test what definition of blend they had used. A formal definition of strength was not given, but all subjects were asked at the end of the test what definition of strength they had used in the test, and what guided their preference choice.

The four samples synthesised for this initial experiment have been described in section 3.2. There should not be a large number of different samples in remote testing over the Internet for the same reasons that samples should not be too long, the main reason being the length of time taken by the listener to download the samples.

Conventionally, tests should be presented in a random order to minimise the effect of weighting caused by having them in a fixed order. However, with such a small number of examples (for four samples to be fully compared, six tests of pairs are required) it was more important to avoid weighting due to the same sample being played simultaneously, so the tests were carefully arranged in a fixed order for all participants. This order combined the criterion of avoiding repeated play of the same sample with an attempt to ensure that each sample was played as both the first and second of a pair in equal amounts, so that weighting due to position in a pair was cancelled out. As each sample was used three times, precise implementation of all the anti-weighting criteria was not possible, but the resulting test order was:

Dull specification – Revoiced specification
Original specification – Bright specification
Revoiced specification – Original specification
Bright specification – Dull specification
Revoiced specification – Bright specification
Dull specification – Original specification

The samples were recorded directly from the organ onto a computer, using the audio-editing program Goldwave (version 4.19) via the analogue audio inputs of a Soundblaster Live Platinum soundcard. The musical example chosen as a typical example of what these ensembles were designed for was a single hymn stanza from the tune "Old 100th". To ensure that each example was musically identical, the hymn stanza was first recorded as a MIDI file (in Cubasis VST version 3.7). For each example, the specification was loaded to the Wyvern electronic organ from the computer via a custom connector and cable. The same combination of stops (as described in section 3.1.1) were manually selected, and the MIDI file was then played back via the MIDI input of the organ and recorded on computer. The basic reverberation unit on board the Wyvern, which is based on an Alesis chipset, provided reverberation for this test. Reverberation presets are not labelled, but the unit was set to Type 4, with 50% depth.

3.6 Results of the preliminary experiments

The results from the preliminary experiment are presented in raw form in Appendix B, and summarised here where appropriate.

The listeners' responses for each of the scales used in this experiment were on a scale of eleven points. The web form gave these answers in numerical form from 1 to 11, as use of negative numbers could have been vulnerable to encoding errors. Six on this scale would represent "neither" or both examples having the same amount of blend or strength. For initial data analysis, answers from all the subjects were averaged, and a standard deviation was calculated to indicate the spread of results.

To look at the overall relationship between the four samples in the test, all listeners' answers for each pair of organs were summed to give an overall result for each pair of samples compared. The summed results then had the null result deducted to remove the offset caused by using the numbers 1 to 11. The null result is the number that would have resulted if everyone had voted "neither". Thus for each pair of samples, the following arithmetic was performed:

$$Verdict = \left(\sum_{1}^{n} answer_n \right) - 6n$$

where *n* is the number of subjects. A positive result of this arithmetic indicated the amount of bias toward the second organ of the pair, and a negative result indicated the amount of bias toward the first organ.

To give each organ an individual rating, the results for each comparison test it was part of were added or subtracted from a running total according to their position in the test. Thus to obtain a rating for the "dull" specification in the test described on page 59, the following arithmetic was performed:

$$Rating_{DULL} = \left(Verdict_4 - Verdict_1 - Verdict_6\right)$$

For the "original" specification, the arithmetic would be:

$$Rating_{ORIGINAL} = \left(Verdict_3 + Verdict_6 - Verdict_2\right)$$

As this result is still expressed in terms of a five point scale, the results are then converted into a percentage of the maximum possible bias by multiplying the rating by 20. This maximum bias would have required all subjects to agree and use the extremes of the rating scales, so is very unlikely to occur in practice, but gives a more readily comprehensible unit of comparison. As results are added to one of a pair and subtracted from the other, verification of mathematical accuracy at this stage is easy by adding together all four results for each organ in each category. If the arithmetical operations have been carried out correctly, all the results should sum to zero, whether expressed as percentages of maximum bias or in raw data form.

Standard deviations are a statistical indication of the spread of results for each individual test, and can also be expressed in terms of the same percentage of maximum spread by multiplying by 20. The formula used here and throughout this thesis is:

$$sd = \sqrt{\frac{\sum_{i=1}^{i=n}(x_i - \bar{x})}{(n-1)}}$$

where n is the population size, and $\bar{x}$ is the average value of x for all n samples. Most of the standard deviations were relatively high, indicating a wide spread of opinion. There were some tests for which there was more agreement, however. To put the standard deviations in perspective, a pseudo-random assortment of fifteen numbers between 1 and 11 (namely 1,2,2,3,4,5,5,6,7,7,8,9,10,10,11) with a mean of six has a standard deviation of 3.21 (64.2%). A collection of fifteen numbers at the extremes and centre of the scale with a mean of six (namely 1,1,1,1,1,6,6,6,6,6,11,11,11,11,11) has a standard deviation of 4.23 (84.6%). Comparison with the standard deviations of the results shows that many are around the 3.2 mark, but the greatest is 3.46 (69.2%).

3.6.1 Preference test results

The preference test results are probably the least useful of the three, but at least give an indication of whether the aims of the initial revoicing were met.

Specification	Result
Default	-3.53%
Revoiced	10.67%
Dull	9.73%
Bright	-16.87%

Table 3.4 – Preference Test Results

The spread of results was rather large, with an average standard deviation for the six preference tests of 2.79 (55.8%). There was most consistency (with a standard deviation of 1.6 or 32%) in the results from the test that compared the bright and dull examples. Many people commented that in this test they could not tell these two apart. Yet the test results show that when compared to the other examples, there is a clear difference between the two samples. The numerical difference in preference between those examples derives mostly from their comparison with the other ensembles. As can be seen from figures 3.1 to 3.4, the bright and dull examples are identical apart from the upper octave harmonics. It should be noted that this and all other overall results in this section fall well below levels of statistical significance, due to the large standard deviations. Statistical significance is a measure that attempts to place results in the context of a null hypothesis. If results diverge from the null hypothesis and can be found in the extreme 5% of the statistical distribution for that null hypothesis, they are said to be significant at the 5% level. Other levels exist, with a smaller level being more significant. Distributions with large standard deviations such as those in this chapter are unlikely to fall in this significant region, but useful trends can be identified and it is important not to solely rely on measures of statistical significance.

Many people indicated in the "Comments" section of the test that their preference was for the brighter ensemble (either "default" or "revoiced"). However, not everyone who said they preferred the brighter ensembles chose those specifications with the most harmonic development. Other listeners indicated that their preference was for one of the duller organs, citing "screaming upperwork" as a reason for disliking the brighter examples. These listeners did consistently pick the duller ensembles.

Some subjects clearly used blend and strength in their preference ratings as well:

> "Mainly, it was blend that struck me as a determining factor: the instruments I preferred all sounded well registered for the hymn tune and I could well imagine singing along to something like these timbres."

> "[I preferred] the overall sound where no particular stop(s) dominate. Some of these mixtures were 'heady' and rather too obvious, although they do serve to provide 'strength' I think."

The comment on mixtures must refer to the upperwork, as subjects were informed of the stops used in the test. This does beg the question of whether all subjects read and, perhaps more importantly, took in the information given to them at the start of the test. As there is indication that this might not be the case, future tests must ensure that information necessary to the test is kept to a minimum.

Some subjects commented on individual organs in more detail. The default specification gained most comment, and was described thus:

> "Thin upperwork seems to penetrate the ensemble, rather than colouring and brightening it."

> "[This organ] had what I would call a more 'cathedral-like' timbre and tone."

> "[This organ] seemed clearer and brighter and more evenly textured, but it also felt as though it would pall on the ear."

One subject described all the organs, and introduced an interesting theory.

> "[Default] is unpleasantly 'scratchy' to my ear. The vowel is inconsistent, as if the upperwork was a later addition. It has a nasty edgy sound, a typical 20th century vowel [ae]. It is very neo-classical and shrill.
>
> [Revoiced] is probably by Continental builder? But the [ah] vowel reminds me of Schultz!
>
> [Dull] sounds typically old-English, with an [o] vowel
>
> [Bright] is the best so far in the test! Is this mid-19th century English or a good modern copy? It has an [o] vowel.

The subject went on to describe a little of the theory behind their vowel descriptions:

> "David Kinsela in Sydney evolved this vowel theory, which seems to explain blend as well as any other factor! It is usually a good indication of the voicing style/period, too (though one can be misled by good modern copies!)"

This vowel theory may be worth examination, as a vowel sound must be related to the formants of the pipe organs in question. Others who have suggested that formants may be significant in our perception of musical timbre are summarised in Bregman (1990, pp. 482-483). However, this is outside the scope of this thesis as developed in chapter one, and while it is easy to become distracted from the core focus of work by interesting tangents, this must be avoided wherever possible.

Overall, the preference tests seem to indicate that the revoiced specification was slightly preferred. However, if the groups are divided by age, there are some interesting results from the preference test. Although subjects were given four age groups to choose from (0-25, 26-50, 51-75 and 75+), to maintain any statistical significance they must be combined into larger groups. When the younger two age groups are combined into one and the older two age groups likewise, this results in groups of nine and six subjects respectively. The following results are obtained for the preference test.

Specification	Result
Default	-12.27%
Revoiced	6.73%
Dull	26.67%
Bright	-21.31%

Table 3.5 – Preference test results for older subjects only

Specification	Result
Default	19.93%
Revoiced	13.40%
Dull	-1.53%
Bright	-31.80%

Table 3.6 – Preference test results for younger subjects only

Both subject groups dislike the bright sample. The older listeners preferred the dull sample, whereas the younger listeners preferred the default sample. Both put the revoiced sample in second place. When comparing the bright and dull samples directly, the younger subject group indicated a slight preference, whereas the older subject group had no average preference for either. Again, the marked difference between the two comes from comparison with the other samples in the test.

Examining the average of the standard deviations of all six comparison tests can give evidence as to how much subjects agree with each other. For all fifteen subjects, the average standard deviation was 2.79 (55.8%). For the older six subjects, the average standard deviation was 2.65 (53%). For the younger nine subjects, the average standard deviation was 2.59 (51.8%). Overall, the standard deviation is less than that of the pseudo-random result derived previously (3.21, or 64.2%), so there is some greater agreement, but still quite a lot of variation of opinion. This should not be surprising as different organists have different tastes and this is a rating scale that is never going to have a precise scientific definition. The younger group, particularly considering their larger size, appears to have slightly more agreement as to their preference.

We have no means of knowing at this stage what might be behind the difference in preference. Conjecturally it could be due to the different kinds of organs subjects have been exposed to, as older listeners will have been introduced to more Romantic-style organs with less prominent upperwork. It could also be related to the ability of listeners to hear higher harmonics, although none of the samples here had particularly strong high harmonics when compared with real-world organ samples.

It is important not to draw too much from this, particularly given the small size of the subject groups. Equally different results might be obtained if, for example, subjects were divided by gender or nationality, but in both cases we lack sufficient subjects in all categories to do this. It will be interesting to see if this apparent age-related difference is evident for the other rating scales.

3.6.2 Blend test results

Specification	Result
Default	-25.8%
Revoiced	2.67%
Dull	16.93%
Bright	6.2%

Table 3.7 – Blend test results for all subjects

The overall results for blend demonstrate some interesting properties. The sample that scored highest for blend was the dull ensemble, as might be expected. This sample was created by altering the default specification in a formulaic manner in line with Sandell's theories of blend. The revoiced sample, altered by ear, scores less well for blend than the bright example. The bright example was created using the dull example as a starting point, so this decrease in perceived blend quantifies the effect of increasing the amplitude of the upper octave harmonics. Overall, these results seem to support the theories of blend discussed in section 3.1.

When the subjects are divided by age, there is less difference between the two groups than for the preference ratings. As these results were all part of one test, the subject groups is the same, with six listeners falling in the older age group category and nine in the younger.

Specification	Result
Default	-40.07%
Revoiced	5.6%
Dull	28.87%
Bright	5.6%

Table 3.8 – Blend test results for older subjects only

Specification	Result
Default	-16.33%
Revoiced	0.73%
Dull	8.93%
Bright	6.67%

Table 3.9 – Blend test results for younger subjects only

The older subjects are more united in perceiving the default specification as poorly blended and the dull specification as better blended. The younger subjects have ratings that are overall on similar lines but to a lesser extent, particularly perceiving less of a difference between the dull and bright examples. The younger subjects also had a higher average standard deviation in this test of 2.89 (57.8%), compared with the older subjects' 2.61 (52.2%).

As introduced in section 3.1, the subjects were also split into two groups, one of whom was given a definition of blend while the other was not. Although equal numbers of volunteers were asked to take each test (about twelve each), because of the problems in testing procedure described in section 3.3 and 3.4.1, only fifteen subjects in total took the test, and these were not evenly distributed between the two tests. Nine subjects had a definition of blend, and six did not. As can be seen from the table of results (appendix B) each subject group had three members who were in the older two age categories, the balance being made up of younger subjects.

Specification	Result
Default	-21.53%
Revoiced	1.47%
Dull	17.8%
Bright	2.27%

Table 3.10 – Blend test results for subjects given a definition of blend

Specification	Result
Default	-32.2%
Revoiced	4.47%
Dull	15.53%
Bright	12.2%

Table 3.11 – Blend test results for subjects not given a definition of blend

Both tables broadly follow the combined results. The main difference is in their comparison of the dull and bright specifications. Here the subjects with a definition of blend appear to have been able to use this definition to help discern between the two specifications. It is important not to make too much of this though, as examination of the raw results show that in direct comparison, six of the nine subjects given a definition of blend did not think either sample was better blended. One of the other subjects was one who always used the extremes of the scales provided, rating all samples either 1 or 11, and thus possibly skewing this otherwise mid-result. Interestingly one other subject in this subset who used the extremes of the scale in the other five comparison tests rated all categories "neither" in the direct comparison between bright and dull samples.

The average standard deviation in the blend ratings for those without a definition of blend was 2.18 (43.6%). For the subject group with a definition of blend, the average standard deviation was 3.14 (62.8%). Far from encouraging subjects to all give the same answer, providing a definition of blend has resulted in a wider spread of results. Providing that all subjects in this subset read and understood the definition of blend, this would seem to suggest that a common understanding of this phenomenon does not equate to a common perception of it when presented aurally. This is an important distinction, and it should be pointed out at this stage that one of the goals in this thesis is to demonstrate the presence or absence of common perception of sound qualities presented aurally, rather than agreement on a particular verbal definition of that quality. Any verbal definition would be derived from acoustic analysis, not human reasoning.

Did the subjects who were given a definition of blend actually read and understand it? In all cases where subjects referred to the definition in their subsequent comments, they claimed to. At the end of the test, subjects were asked to comment on whether they agreed with the definition of blend given at the beginning of the test. To recap, this definition of blend was "the tendency for concurrently sounding timbres (for example the many pipes in a pipe organ) to fuse into a single timbre".

Five subjects stated that they agreed with the definition, and two made no direct answer to the question. Two subjects stated that they did not agree with the definition, and went on to describe why:

> "I *used* the above definition - but I don't agree that a good blend should be bland! Total fusing for me makes it bland. It's more to do with the 'mixing' of the overall sound. So I think I like a good MIX, not a good BLEND"

> "No. I think the definition of blend should include the awareness and appreciation of different voices, as long as one does not overpower ("drown out") another and there is clarity of sound (no muffled, "muddy" sounds). Based on the way I was taught, no one stop should be added to an ensemble just to have more stops engaged. Such a practice may destroy the effectiveness of already blended voices. At the risk of sounding bromidic, sometimes less IS more, because of the clarity provided by a few simple, but lovely stops."

Both comments seem to be objecting to the implied homogeneity of a blended sound. Both also have assumed that blend, whatever its definition, should always be a desirable quality, whereas for the reasons elucidated in these comments, total blend is not always a good thing in an organ ensemble or indeed any other musical ensemble.

Of the subjects who agreed with the definition, many went on to comment further:

> "I would agree; it's what I look for in an ensemble, where no single voice stands out soloistically over the other voices (except when done intentionally). A chorus at 8.4.2-2/3.2 should sound as one voice. If a single register is overbearing on one note, the chances are that the

> voicing or finishing of the entire chorus will be off and there will be other voices standing out on other notes"

> "Blend to me is fullness of tone, absence of uncertainty of the mixture."

Many commented that blend was not something they specifically thought about, but was a quality inherent to a good combination of organ stops. Some subjects gave vague answers that demonstrated a lack of ability to define blend in terms of words.

People who were not given the definition of blend were also asked at the end of the test what definition of blend they used. It is interesting to compare these with the definition of blend and comments from listeners described above. Their answers included:

> "How well the individual stops blended into the chorus."

> "A pleasing mix of harmonics low to high. 'Pleasing' may be subjective"

> "Blending is balance in high and low pitches"

> "Ranks balanced in loudness, and most important, they must all have a consistent vowel formant (often a 'give-away' when ranks come from different builders or periods or are revoiced, especially reeds!)"

> "The ensemble should be cohesive; no single voice should predominate"

> "An overall sound where no particular stop dominates, where there is a sense of cohesion about the sound."

All these answers seem to broadly concur with the theories of blend described previously. However, what is clear from the high standard deviation of the blend ratings for all subjects is that people are less consistent in how they apply their judgement of blend to the ensembles they hear. Agreement may be greater where the samples used are not so similar.

3.6.3 Strength test results

Specification	Result
Default	13.27%
Revoiced	8.47%
Dull	-9.27%
Bright	-12.47%

Table 3.12 – Strength test results for all subjects

Here the average standard deviation is somewhat less than for other tests, at 2.23 (44.6). This suggests somewhat more concurrence on what strength is in the context of pipe organs. The theory of strength described in section 2.2 was that for an organ to be perceived as strong, it should have both solid unison (or sub-unison) foundations combined with a rich harmonic development reinforcing the foundation's harmonics. As all four pipe organ examples here are based on the same foundations, this judging of strength can only verify the second part of this theory. Comparing the results in table 3.12 with the ensemble design and analyses in section 3.2, this theory appears to be supported. Perceived strength decreases as the higher harmonics are removed. When only some of those harmonics are reintroduced (the bright example), this does not add to the perception of strength. This confirms the theory's wording that a rich harmonic development (with no significant harmonic gaps) is necessary for the perception of strength.

However, these overall results hide distinct disagreements between the older and younger subjects. Using the same age groupings as before, the following results are obtained.

Specification	Result
Default	-12.2%
Revoiced	6.67%
Dull	5.53%
Bright	0%

Table 3.13 – Strength test results for older subjects only

Specification	Result
Default	30.4%
Revoiced	9.6%
Dull	-19.27%
Bright	-20.73%

Table 3.14 – Strength test results for younger subjects only

Here the younger subjects distinguish more clearly between the examples, supporting the theory to a greater degree. The older subjects found little difference between the four examples, and with an average standard deviation of 1.68 (33.6%) were broadly in agreement on this. Perhaps the most significant result is that the older subjects believe the default example to have the least strength, whereas the younger subjects rate it as having the highest strength.

Subject comments may provide insight into this disagreement, and are presented here with their associated age groups to see if this is the case. Listeners were asked the question "What definition of strength did you use, and/or what do you think strength is in this context?" Answers from both age groups included many on a common theme:

> "Strength to me is the overall 'weight' of the organ sound." (older)
>
> "Some sounded louder because they sounded brighter i.e. with more high tones." (older)
>
> "Whether one organ sounded louder" (younger)
>
> "Apparent loudness - difficult to define." (older)
>
> "Volume" (younger)

Subjects from both age groups appear to relate strength, described above, to perceived volume. Hall (1993) describes the concept of perceived volume and its relationship to the harmonic development of the sound in more detail. The sounds perceived as stronger by the younger age group do have greater harmonic development in a manner that might suggest that they were being played louder.

There was a second common answer given, perhaps triggered by the use of a hymn fragment as the musical example:

> "A 'stronger' sound can be heard more clearly, and would support (and encourage) singing with more volume and spirit." (younger)

> "Solid foundation for leading a hymn. Clear tone that can easily be followed" (older)

> "A feeling as to whether it could be heard over a congregation singing in terms of leading that singing."(younger)

> "Strength for me is the ability to clearly support the congregation I serve as they sing" (younger)

Strength is clearly important to these subjects in leading hymn singing, and this may be the primary function of the organ in many churches. Other respondents seemed to go against or elaborate upon the more common answers, or were less clear in their answers indicating the need for better understanding of all the timbral semantics they used:

> "I felt the brighter, top-heavy tones to be stronger simply because those tones tend to excite the nerves and engage the hearing more than do 'unison' and other tones below the top-most pitch levels." (younger)

> "Strength on an organ to me would be that it has bite without becoming too muddy. Just because we can open [up] an organ and play it loud doesn't necessarily make it the stronger organ." (younger)

> "To me, strength comes from the breadth of the unison stops. A broad, rich, and warm Principal will provide a very solid foundation on which the octaves and mutations can rest. It has always amazed me how much strength (but not loudness) can be achieved by adding a large Open Diapason or a Gross Gamba to the ensemble, rather than adding a mixture!" (older)

> "Strength is to me the prominence of certain timbres as a part of the whole texture." (younger)

> "I assumed here that you meant 'sharpness', which tends to mean 'bolder' rather than just louder." (older)

There doesn't seem to be any clear distinction between the age groups in terms of their answers to this strength question that would give clues to their differing perception of strength. The theory presented appears to be well supported by younger subjects but not by older ones, who distinguish rather less between the four samples for this rating scale.

3.7 Summary of initial experiment and related issues

Subjects were united in their dislike of the bright example, but differed with age according to their preferred alternative. The two age groups agreed most on which were the better blended examples, but disagreed on which were the strongest.

In the case of perceived blend, both subsets of subjects (those who were given a definition of blend and those who were not) had a large standard deviation, and giving a definition of blend if anything slightly increased that disagreement. The definitions of blend given by those who were not given one at the beginning of the test broadly concurred with the definition given to the other subject group. This suggests that the concept of a blended sound is commonly understood, but that this does not translate into a common understanding in the psychoacoustic domain. Strength lacked consistent understanding overall, although the younger age group had a distinct common understanding. This did not translate into a specific common verbal definition of strength, although the definitions given mostly fell into two categories and all appeared to be in the same broad area. Here again the definition of a term appears not to improve actual psychoacoustic agreement, although subjects appeared to have a common linguistic understanding of the term.

This initial experiment had many areas that could be improved upon. The small numbers taking part means that detailed analysis of age-related differences may tend towards examination of experimental noise. It may be the case that the preference for certain examples is only valid within this group. Subjects might normally have other criteria on which they rate pipe organs, but in this instance have fallen back on simple criteria, as all the examples sounded similar. The four samples were created with small, quantifiable alterations from a single starting point. In this attempt to reduce the differences between the four samples to known parameters, the samples were so similar that some subjects were not able to tell all of them apart. It may be that agreement on rating scales is greater where the samples used are not so similar.

However some useful results have come out of this exploratory stage of testing, as well as giving ideas for improvements in test procedure for subsequent experiments. Future tests must have many more subjects until the nature of common understanding becomes clear. Many of those taking the test should also do so in person (as opposed to over the Internet) as a control group against any effects of the Internet hereto unidentified.

Two important areas to consider are the statistical analysis of how test procedure affected the results, and whether the results are inter-related. Such correlation could either be due to actual psychoacoustic inter-relation, or test procedure implying a relationship between the parameters to the subjects.

3.7.1 Consideration of skew in test results

The issue of test procedure can be examined by looking at the mean of the results in each category. Ideally, in the case where each sample occurs in its pair first and second in equal amounts, the mean of all results would be approximately the "neither" result: six, in this case. Significant deviation from this mean, particularly with large numbers of subjects, would suggest that the results had been skewed by the test procedure and one sample had been given undue prominence by its position. Unfortunately the samples in this case are not equally distributed in first and second, as this is impossible with only four samples and a minimum number of comparison tests. Evidence of skew may only point to preference of the sample or samples that occurred twice on one side. However, as each sample occurs twice in one position and once in the other, any deviation of the mean outside the central third would point to a skew resulting from test procedure and not sample preference. The only way to definitely determine whether any skew resulted from the order in which samples were placed would be to repeat the test with another group of subjects and the sample order reversed. As this is impractical here, future tests using this procedure should also look at the mean of results to isolate any skew due to testing procedure.

The overall means of the tests in this experiment are presented in table 3.15

Test name	Mean test result
Preference (all subjects)	5.51
Preference (older subjects)	5.19
Preference (younger subjects)	5.72
Blend (all subjects)	5.67
Blend (older subjects)	5.31
Blend (younger subjects)	5.91
Blend (with definition)	5.85
Blend (without definition)	5.39
Strength (all subjects)	6.21
Strength (older subjects)	5.53
Strength (younger subjects)	6.67

Table 3.15 – Mean result for each test category

All mean result values are well within the central third of 4.33 to 7.66, and there is no significant evidence of experimental skew. The older subject group appears to have a slight overall bias towards the first of any pair of results, but this group only contained six members so this could be due to small sample size rather than test procedure. The group who took the test without a definition of blend also contained six subjects. It would be worthwhile to review the possibility of experimental skew in future tests, particularly with regard to older subjects. However at this stage the slight variations either side of the central value in mean results are easily explained by the unequal test distribution, and present no cause for concern as they are very far from statistical significance.

3.7.2 Consideration of correlation between test results

A second useful consideration related to test procedure is to consider whether there is an inter-relationship between the three ratings scales in the experiment. Any such relationships discovered may not be due to test procedure but could be an interesting secondary experimental result. However, if all subjects' answers are very closely related, the possibility of a suggested link due to experimental procedure must be examined. In this particular experiment, it will be interesting to see whether subjects' preference is related to either perceived blend or perceived strength. Of more importance from a procedural point of view is any connection between blend and strength. As these samples have been artificially created based on two separate theories, such a connection could either indicate a problem with experimental procedure, or cast doubt on the accuracy of any theories derived from study of individual results without reference to other comparison scales.

Several statistical methods exist for the testing of correlation between two sets of results, one of the most common of which is the Pearson correlation coefficient, which is the most useful in this instance. The formula for the Pearson coefficient is:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

where x and y are the two sample sets, $\bar{x}$ and $\bar{y}$ their respective means, and r is the correlation coefficient. r can vary between +1 and -1. Zero represents no correlation, +1 represents the highest possible positive correlation, and -1 indicates the highest possible negative correlation, which is where a positive alteration in one variable correlates with a negative alteration in the other variable. The output value varies in statistical significance according to the sample size, and tables of values are produced which allow a particular result to be assessed at various levels of statistical significance. Further details of this method, significance testing and the tables used to rate significance can be found in Howitt and Cramer (2000).

This method only works with two sets of variables, so the correlation between each pair of rating scales must be considered individually. Measuring correlation in this manner can have problems where there is a single anomalous result, as the formula cannot adequately isolate that result. Examination of the raw test data in appendix B shows one subject who used only the extremes of the scales. The covariance results with and without his sample included are presented in table 3.16.

It is dangerous to disregard exceptional results simply because they do not fit the general trend. In a more traditional testing environment, it may be that the subject misunderstood experimental instructions. However, in this case the subject was one of those from the USA who took the test over the Internet, and it is therefore impossible to study why his answers varied from the others in the same detail. Results are therefore presented with and without this subject's answers so that both the general trend and the complete body of results can be examined.

For a relationship between two variables to be statistically significant in this instance, its Pearson correlation coefficient should consistently exceed the minimum significant value for the sample size over all six tests. For the test with 15 subjects, a value of less than -0.51 or greater than +0.51 is significant at the 5% level. Where the exceptional result is excluded, this increases the statistically significant value to -0.53 and +0.53 respectively.

All results	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Blend/Strength	0.23	-0.34	0.23	0.56	0.14	0.35
Blend/Preference	0.55	0.62	0.59	0.74	0.43	0.65
Strength/Preference	0.67	0.28	0.77	0.87	0.69	0.71
Excluding exceptional result						
Blend/Strength	0.50	-0.21	0.53	0.56	0.54	0.62
Blend/Preference	0.92	0.83	0.95	0.75	0.87	0.97
Strength/Preference	0.58	0.13	0.68	0.87	0.61	0.62

Table 3.16 – Covariance results for each test

Where both strength and blend are compared to listener preference, in all but one of the tests the correlation is statistically significant. When the exceptional result is discounted, the level of correlation between blend and preference becomes strongly significant. One test remains insignificant in the comparison between strength and preference, and overall, the level of correlation between these two parameters is significant but less so than for blend and preference.

The insignificant test is the comparison between the original specification and the bright specification, which had fairly high standard deviations for both preference and strength. As might be expected, the correlation between blend and strength is also exceptionally low for this test, hovering around the level of significance for all other tests.

Separating the subjects into age groups as has been done in previous analyses reveals some other interesting trends. The exceptional result has been excluded, resulting in an older age group of six subjects and a younger age group of eight. Statistical significance for a group of six subjects is below -0.81 or above +0.81, and for a group of eight it is below -0.71 or above +0.71.

Younger subjects only	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Blend/Strength	0.40	-0.74	0.53	0.74	0.47	0.74
Blend/Preference	0.95	0.80	0.94	0.71	0.78	0.95
Strength/Preference	0.29	-0.56	0.74	0.98	0.61	0.76
Older subjects only						
Blend/Strength	0.37	0.59	0.42	0.46	0.68	0.30
Blend/Preference	0.87	1.00	0.97	0.96	0.99	1.00
Strength/Preference	0.67	0.62	0.58	0.51	0.70	0.30

Table 3.17 – Covariance results for each test divided by age

Both groups retain consistent and significant correlation between their ratings for blend and preference. In the case of the older subject group, this correlation is very strongly significant. Although general trends of correlation can be identified in other parameter pairs, these do not consistently fall around or above the levels needed for statistical significance.

It can therefore be concluded that a well-blended ensemble is a key factor in listeners' preference for one organ over another, and that this is particularly the case in older listeners. Other correlations are less significant, but the fact that there are some consistent trends suggests that this is a useful source of additional information on the inter-relationship of various timbral descriptors.

The question of whether this correlation is due to test procedure must now be answered. As correlation varies between results, overall the test does not appear to have implied a relationship. Strength has neither a significant positive or negative correlation with blend, which suggests that the subjects were not treating the two scales as related either directly or as opposites. The fact that different subject subgroups have differing levels of correlation between the parameters could either suggest that the test procedure was affecting subjects differently according to their age, or that test procedure was not affecting subjects' natural and differing answers. The latter seems more likely than the former given the relatively simple test procedure.

3.7.3 Other considerations for future experiments

Dividing subjects by age has produced some significantly different results, and therefore it is important to continue to study this in future experiments. It would also be interesting to consider factors such as country, mother tongue, sex, musical preference or playing ability. In all these cases, this experiment had too few subjects outside of the majority category to make these analyses worthwhile, and it may be that future experiments suffer from the same problem for some factors. However if this data is gathered, the option of future analysis remains open.

As there were only four listeners who rated their playing ability as less than good, any comparison of results according to playing ability would be flawed. However the question of ability is an important one, and future experiments must return to this question to determine whether it has a significant impact on listener ratings. In general, all listeners involved in this initial experiment concurred on some important factors, with the exception of one listener whose self-rating for playing ability was good. This suggests that by inviting listeners from a certain field of interest, even if those listeners themselves are not musical in terms of playing ability, their involvement with the pipe organ and volunteering for the test could be a form of self-selection of appropriate candidates.

Several subjects spotted that the examples and/or the reverberation were artificial. If these and future results are to be valid in the timbre space of the pipe organ, it is important to make the examples as realistic as possible. In the case of synthesised examples, improvements must be made in both these areas. It may be appropriate in some tests to use real examples, and as this is more time efficient compared to the complexity of synthesising better artificial examples, this should be done wherever it is appropriate in preference to the use of artificial examples.

Many subjects commented on one common problem with the tests. The samples seemed to jump on initial playback. The cause of this was again an "update" in software, this time caused by the browser "plug-in" which handles sound playback. Previously, samples had been downloaded fully before being played back. However, the more recent "QuickTime" plug-in aggressively takes over the playback of most sound samples, and attempts to play them back as it is downloading them. Even when this is done from a local hard disk, the result is interference in the first second of playback. If future tests are conducted using the Internet, alternative sound playback tools should be investigated that do not attempt to play the sample before it is fully downloaded. One possibility is to "embed" the sounds in the test web page, rather than linking to separate sound files. This solves the problem without altering subjects' computer setups, but not all computers are able to make use of this function, so the existing method must also be available as a backup.

Subjects also suggested the use of alternative tunes, and it is worthwhile considering alternatives that might better demonstrate the whole ensemble. In ensemble playing, there may be a difference in perceived qualities of the ensemble between those pieces that specifically use polyphony such as counterpoint, and those for whom it is more chordal. Certainly the work in Disley (2000) summarised in section 2.2 suggests this might be the case. However, it is important to focus on the primary purpose of most pipe organs, which is to provide music of a more chordal nature in hymns.

One subject asked for the tune to be played slower so as to better appreciate the ensemble. Speed of playback is a factor in reducing the download time, but future examples must not be played too fast to allow each note to sound fully. One subject also commented: "Now I'll never get that tune out of my head" which suggests the onset of experimental fatigue and memory effects across examples. This indicates that to lengthen the tests, perhaps in pursuit of an equal balance of samples to avoid experimental bias, would have unfortunate consequences. Subjects would become increasingly fatigued by repetition of the same musical example and their answers might become less accurate as the test progressed.

3.8 Conclusion

This chapter has introduced a number of concepts of experimental procedure and analysis of results while describing an initial experiment. That experiment has demonstrated that there is common understanding for some timbral descriptors across the subject groups. The way in which that understanding varies across subsets of subjects has suggested that future experiments should consider various factors such as age when analysing the results. Various alterations to technique for both experimental procedure and results analysis have been suggested in light of this experiment, but overall the procedure appears to be sound and does not introduce significant bias into the results.

Several other interesting phenomena have been observed, such as the strongly significant correlation between perceived blend and preference for older subjects. In conclusion, the initial experiment suggests that, with slight modifications, the experimental procedure proposed is a valid methodology for the investigation of a hypothesis that appears to be supported by initial results.

4. Word Gathering

This chapter considers the research strategy introduced in section 1.6 in light of the conclusions of chapter three. A means of gathering adjectives is developed and the initial stage of gathering timbral adjectives is undertaken in section 4.2. Section 4.3 considers listener classification of those words, and section 4.4 describes the process of selecting words worthy of further study.

4.1 Development of research strategy

In section 1.6, the research strategy was described thus:

> The proposed research will commence by study of existing timbral semantic work, and look to verify those findings within the timbre space of the pipe organ. Analysis of the results of those initial experiments should then fine-tune the next stage of work, which will look to gather timbral adjectives from listeners in a non-biased manner. The most common of those adjectives will then be used as comparative ratings scales in listening tests.

Chapter two has introduced existing work on timbral semantics and related areas, and chapter three has examined some of the author's theories developed in earlier work within the timbre space of the pipe organ. Several things have become clear as a result of that work, which will influence the way the remainder of this body of work is conducted.

A common understanding of the verbal definition for a term does not translate into a common application of that term. Defining adjectives is therefore unhelpful, quite apart from the obvious concerns that doing so biases answers towards a particular conclusion. Samples should be as realistic as possible, and possibly vary more than in the initial experiment so that subjects can perceive a difference between all of them. Finally, other subject factors such as their country, sex and age should be considered when analysing results.

The goal of the research in this chapter is to gather timbral adjectives from listeners in a non-biased manner. These adjectives have to be related to the timbre space of the pipe organ.

The timbre space of the pipe organ ranges from the traditional principal ensemble, through choruses of flutes to quiet string stops. Reeds can either be imitative solo stops or harmonically enriching chorus stops. These broad categories are recognised by organists and each has a different role. While all fall under the broad timbre space of the pipe organ, it is sensible to compare the attempts of different organs to achieve the same function. The author discovered in previous research (Disley 2000) that the introduction of chorus reeds in some samples made it difficult to isolate the psychoacoustic triggers for certain words. Of all the ensembles, it is the principal chorus that is most significant in the organ's role as an instrument that both accompanies singing and performs solo literature. This study will therefore constrain itself to the principal chorus in its varied forms. The research described in this and subsequent chapters was summarised in Disley and Howard (2003), although there are a few minor differences between that paper and this thesis, largely due to further analysis.

4.2 Procedures and results of previous studies

In previous research on this subject, adjectives have either been picked arbitrarily by the authors, or gathered from a limited number of listeners by directly asking them for a list of adjectives they used. The first method is similar to that used in Disley (1999 and 2000), where of a small number of terms used by subjects, two ("blend" and "strength") were chosen for further research based more on interest than measured frequency of occurrence. The second method was used by Moravec and Štěpánek (2003), who asked 120 musical subjects for the words they used to describe timbre. Subjects were players of a number of different instruments, and their results were analysed by class of instrument. This method was also used by Rioux (2000) who used just one specialised subject (an organ-builder) as a source of adjectives.

Both methodologies have flaws. If the adjectives are picked arbitrarily by the authors, they may not be those used commonly by people to describe pipe organs. However, if subjects are asked directly for the words they use, they will come up with a list of words some of which they may not use in practice, and with no indication of which words they would more commonly use. A solution to this is not to ask the subjects for words, but to play them a number of different musical examples and ask them to describe the sound of the organs they are hearing.

It seemed sensible to proceed with the indirect audio-based method of adjective gathering, with the possibility of more direct questioning if response is low. This was not required in eventuality.

It is interesting to compare this methodology with two parallel research projects. Rioux and Västfäll (2001, a & b) conducted a similar experiment to the one described in this chapter, but using a very limited set of recordings of single pipes of one pitch and type at different stages of voicing. Nykänen and Johansson (2003) used a combination of directly asking subjects for words and asking subjects to describe recordings of saxophones. Additionally, they asked the ten saxophone playing subjects to describe the sound they were aiming to create when they played.

As these projects were being carried out concurrently, their methodology was not known to the author of this thesis at the time his experiments were being designed, but they have similar problems to contend with. It is interesting in particular that Rioux and Västfäll tackled a related experiment with a very different goal in a similar manner.

4.3 Experimental procedure to gather adjectives

The aim of this part of the experiment is to gather as many adjectives as possible from as large a subject group as possible. It is anticipated that some of those adjectives will be more common than others, and therefore worthy of further study. The use of an indirect test method, namely listener response to different audio samples, will hopefully exclude those words that seem intellectually appropriate but which are not used in practice.

The desire to involve as large a subject group as possible, coupled with the requirement of prior knowledge of the timbre space of the pipe organ, make continued use of the

Internet as a test medium an obvious solution. This does not preclude the direct involvement of subjects who can take the test in a controlled manner at the author's institution. However the goal at this stage is only to restrict the timbre space to that of the pipe organ. Comparison tests would be inappropriate at this stage, for if samples are limited in number to avoid subject fatigue, the tests would have be within a subset of pipe organ timbre space so that subjects did not feel they were comparing two unrelated ensembles. The use of such tests will be highly valuable in examining consistent understanding of the most common words at a later stage in this experiment.

4.3.1 Recording and choice of samples for experimental use

Previous research in this area has largely been limited to the sound of single pipes, so the sample libraries used are not necessarily appropriate for this experiment. For example, Creasey (1998) used samples of single organ pipes as well as other single note instrumental samples from the MUMS (Magill University Master Samples) sample set. There are no equivalent standard libraries of pipe organ timbre. Rioux and Västfäll used recordings they created from multiple copies of a single historical organ pipe.

In the course of his previous research (Disley, 1999 and 2000), the author had built up a collection of recordings from a number of pipe organs, including both single note samples for experimental analysis and collections of musical phrases played on a number of different stop combinations. It seemed to be a prudent use of time and resources to use these at least for the initial adjective gathering experiment, as they had been recorded in a controlled high quality manner. It is not necessary for this part of the experiment for all examples to be the same musical extract. However, playing only one extract eliminates the possibility of different musical extracts producing different emotions (with the potential for indirectly affecting the timbral adjectives) in the listeners. The words produced by this adjective gathering can therefore be said to be descriptions of the pipe organ ensemble sound rather than comments on musical style or emotion.

Previous samples had all been recorded using a Sony ECM-999PR stereo microphone and either a TEAC DA-P20 or Sony TCD-D7 portable DAT recorder set to 44.1kHz at 16 bits resolution. Samples were transferred digitally to a computer to avoid degradation of the audio signal. There were a number of different organs to choose from, but to avoid confusion details will be only given of the organs and stop combinations chosen wherever they are used. Details and specifications of all organs used throughout this thesis can be found in appendix E, to allow the reader to put the choice of organs and registration in context. The sample library includes a variety of British organs with a variety of historical styles represented.

The samples had been recorded with the microphone placed approximately 20 to 25 feet away from the pipe organ, in line with the advice given in Horning (1998). The way in which the sound of the pipe organ varies with distance from pipes is explored in Syrový et al. (2003). Each position in the room will potentially experience a different sound from the pipe organ, but within the context of this experiment (and indeed the previous ones for which the samples had been recorded) it is sensible to restrict potential microphone positions to potential listener positions. When a pipe organ is voiced, the voicer is conscious of the effect of the room acoustic on the sound delivered to an audience in the room, and will adjust the tonal balance of the ensemble to sound at its best for audience listening positions. Within that subset of potential microphone positions, Horning suggests that the distance of 20 to 25 feet best combines clarity of the source with sufficient room acoustic to place the organ in its auditory context.

Many researchers use a dummy head in their research to take the simulation of a listener position in the room one stage further. However, to make the best of this highly accurate method of recording, casts of listeners' own ears should be used on the dummy head, or at least the nearest equivalent from one of six standard sets of ears. The sound should also be delivered to the subjects as close to their own ear canals as possible, ideally using in-ear headphones (Pierce, p100, in Cook 1999). These two requirements make the use of a dummy head difficult with a large group of subjects, particularly where the subjects are geographically dispersed and comparison tests are used to obviate the necessity of standard test equipment. Subjects were asked to use conventional on-ear headphones to remove the effect of their own acoustic environment, although as with all conventional stereo sound reproduction methods, the listener's position in the recorded environment cannot be adjusted. This ensures that all listeners experience the same listening position, removing a possible variable present in real life, although possibly detracting from the desired effect of reality.

From the library of samples available, four were chosen to demonstrate the widest range of ensembles. All four were in environments that had some degree of natural reverberation. The four organs and ensembles used in this test were:

- Heslington Church: Open Diapason, Principal, Twelfth and Fifteenth (Great)
- Jack Lyons Concert Hall: Principal, Octave and Flachflöte (Hauptwerk)
- St. Chad's: Open Diapason, Principal, Fifteenth, Mixture (Great)
- St. Columba's: Open Diapason, Principal (Great), Superoctave (Swell, coupled)

It could have been possible to include instruments or registrations that while within the timbre space of the pipe organ principal ensemble were considered poor (for example out of tune or unbalanced ensembles). However, this might have led to subjects concentrating only on the negative aspects of certain ensembles rather than describing the four very different ensembles chosen, and thus produced a less useful and generally applicable set of words.

One problem was that the musical samples were originally recorded for a CD-based listening test, so were larger than ideal for Internet listening tests. The solution was to restrict the length of the samples. It proved possible to isolate the first phrase of "Old Hundredth" and cut off just before the beginning of the second phrase, preserving some of the reverberation necessary to hear the ensemble in context. The resulting samples were between seven and eight seconds in length, with file sizes of up to 1.4MB. The arguments about how to best reduce file size have been presented in section 3.4, along with the desired maximum file size of 1MB and the reasons why lossy compression is inappropriate. The reasons behind the decision to use mono as opposed to stereo samples in the initial experiment did not apply here. As there was negligible spectral energy above about 10kHz, the decision was made to present the samples in stereo at the expense of a halving in the sample rate.

The samples were therefore presented to the users in uncompressed files of between 628kB and 714kB, sampled at 22.05kHz at a 16-bit resolution. Samples were equalised for pitch utilising the resampling function of Goldwave version 4.19 (an audio editor), which uses an interpolation based method to avoid artefacts that could be introduced by FFT or other pitch alteration methods. Such pitch correction was minor, never exceeding 50 cents.

4.3.2 Test subjects and procedure

Fifty-three subjects took part in the test, well in excess of the suggested minimum numbers in Levetin (p310, in Cook, 1999). A small number of subjects had different problems with remote test procedure, and either were unable to complete the procedure or submitted some answers more than once. Where subjects were not able to complete all four examples, the adjectives they had used were included, as it was not necessary to hear all examples before commenting on any one. Where the same subject used the same word twice as a result of submitting a duplicate or similar answer about the same organ, this was discounted. Where subjects used the same word to describe more than one organ, this has been noted in the table of results (table 4.2 below).

The samples were presented to subjects, both in person and over the Internet, via a web page interface similar to that used in the initial experiments. Some subjects had previously commented that their web browsing software seemed to be unable to play more than one embedded sample, so the test was divided into one page for each sample. This cured the problem, but at the expense of getting incomplete results for some subjects who for various reasons did not complete the entire test. This method is therefore inappropriate for experiments where it is important that subjects complete all tests, such as comparison tests. Subjects may feel that as they have submitted some of the answers, it is less important to complete the test if they are forced to break off. With a single submission test, comments received from volunteers explaining why they have delayed taking the test suggest that they set aside a single block of time to complete the test without interruptions. Such considerations are obviously only relevant to remote subjects either taking such tests with large downloads over a slow Internet connection or taking Internet based tests with a large number of questions.

Geographical data was collected about subjects, and is presented in table 4.1 below. Numbers in the second column are those who did not complete the entire test, but submitted at least one description out of a possible four. As the test was sequential, in all cases these partial answers were for the earlier organs in the sequence.

Country	Subjects who gave full answers	Subjects who gave partial answers
United Kingdom	10	
United States	31	3
Canada	3	
Venezuela	1	
Greece	1	
Australia		1

Table 4.1 – Geographical distribution of test subjects

One subject was Venezuelan, speaking English as a third language after Spanish and German. Another subject was English but living in Greece. One American subject originally spoke German but could no longer do so (although they retained their ability to understand German speech). All other subjects were natives of the country of birth and spoke English as a first language.

Subjects were also asked to rate their musical ability. Due to the problems observed in self-classification in the previous experiment, and the fact that the data being gathered was not of a nature that could easily be analysed alongside such factors as age or playing experience, it was decided to experiment with a description based classification system. Subjects were asked to describe their experience of listening to and playing the organ, and those that could be easily categorised are presented in table 4.2. Because of the nature of the question, the categories are not mutually exclusive, with the exception of the playing ability categories. As not all subjects gave a specific answer to their playing ability, the numbers in those categories will not add up to 100% of participating volunteers.

Experience categories	Subjects who fell in this category
Good playing ability	33
Some playing ability	14
No playing ability	2
Pipe organ builder (professional)	8
Pipe organ builder (amateur or limited experience)	8
Broadcaster or recording engineer	2

Table 4.2 – Pipe organ listening and playing experience of subjects

While interesting, it is impossible to make any meaningful analysis of such results because of a lack of precise grading of subjects' musical abilities. Unless a good method of gathering such data accurately can be found, it seems sensible not to gather it as both methods have produced results of limited use. What this data does show is that in general the volunteers for this test have considerable experience of the pipe organ, either as players, builders or listeners, and are therefore well qualified to comment on its timbre space.

While a random order of tests was not necessary in gathering adjectives, a secondary motive meant that different subjects were given different images of each organ. It was unclear from the author's previous research (Disley, 1999 and 2000) whether a picture of the organ being played should be shown to the subject to enable them to better place the sound in context, or whether the picture shown would influence their description of the organ. When sufficient subjects volunteered, the opportunity was taken to divide them into three subsets.

Each group was presented with the same sound samples, but the order of image presentation was altered. One group was presented with the correct sound examples for each picture, and the other two were altered in such a way that no sample was played with the correct organ. The lack of a fourth group meant that the least likely combinations of organ and sound could be avoided without any combination of picture and sample being presented to more than one subject group. Subjects were particularly asked to describe the sound of the instruments they were listening to, rather than its appearance or the acoustic. It could then be determined whether the presentation of images was unwittingly influencing their description or perception of sound. The results specific to this part of the experiment will be considered in section 4.5

4.4 Results of adjective gathering

Table 4.3 includes every adjective used by the subjects in describing any of the organs. The numbers in brackets after each word represent the frequency of occurrence. The first number excludes repeated use by the same subject when describing different organs, and the second number includes repeated use. So, for example, an adjective with (5/7) after it was used by five different subjects a total of seven times. The most common adjectives are presented in figure 4.1 below. Adjectives were used in both positive and negative ways – most referred to qualities (positive or negative) the ensemble in question had, but some referred to qualities lacking in that particular example. Multiple variants of the same word (e.g. “has character”, “characterless” and “characterful”) have been grouped together. The raw results from which these adjectives have been gathered can be found in appendix C.

Abrasive	(1/1)
Aggressive	(2/2)
Agreeable	(1/1)
Anaemic	(1/1)
Articulate	(7/12)
Authoritative	(1/1)
Balanced	(11/12)
Baroque (neo)	(4/7)
Bland	(1/1)
Blended	(7/9)
Body (has)	(2/2)
Bold	(1/1)
Breathy	(4/4)
Bright	(26/41)
Brilliant	(2/2)
Brittle	(1/1)
Broad	(1/1)
Busy	(1/1)
Buzzy	(2/2)
Calming	(1/1)
Cerebral	(1/1)
Character (has)	(3/4)
Chiffy	(6/7)
Clean	(6/8)
Clear	(15/19)
Close	(1/2)
Cohesive	(3/4)
Comfortable	(1/2)
Commanding	(2/4)
Competent	(1/1)
Complex	(1/1)
Creamy	(1/1)
Crisp	(1/1)
Crunchy	(1/1)
Cutting	(1/1)
Dead	(2/4)
Dignified	(1/1)
Distinct	(2/2)
Driving	(1/2)
Dull	(3/3)
Dynamic	(1/1)
Edgy	(4/4)
Elegant	(2/2)
Engaging	(1/1)
Exciting	(5/6)
Even	(1/1)
Firm	(2/2)
Flat	(1/1)
Fluffy	(1/1)
Flutey	(10/12)
Forced	(1/3)
Full	(9/11)
Full-bodied	(1/1)
Fuzzy	(1/1)
Gentle	(5/5)
Grand	(1/1)
Grating	(1/1)
Hard	(1/1)
Harsh	(3/3)
Heavy	(2/2)
Hollow	(2/2)
Homogenous	(1/1)
Immediate	(1/1)
Integrated	(1/1)
Interesting	(7/7)
Intrusive	(1/1)
Jangly	(1/1)
Lazy	(1/2)
Lean	(1/1)
Light	(7/7)
Lively	(3/4)
Majestic	(2/3)
Mellow	(5/5)
Metallic	(1/2)
Mild	(1/1)
Muddy	(2/2)
Muffled	(1/1)
Nasal	(2/2)
Nice	(11/14)
Open	(3/3)
Overbearing	(2/2)
Piercing	(3/4)
Piquant	(1/1)
Plain	(3/3)
Pleasant	(11/14)
Punchy	(1/1)
Pungent	(1/1)
Pure	(1/1)
Quiet	(1/1)
Reedy	(7/8)
Resonant	(3/3)
Restful	(1/1)
Rich	(7/8)
Robust	(3/3)
Romantic	(2/2)
Rough	(1/1)
Round	(3/3)
Scratchy	(1/1)
Screechy	(1/1)
Shimmery	(1/1)
Shrill	(1/4)
Singing	(4/4)
Silvery	(1/2)
Sizzly	(1/1)
Sluggish	(1/1)
Small	(2/2)
Smooth	(3/3)
Soft	(3/3)
Solid	(4/4)
Sparkling	(2/3)
Spiky	(1/1)
Steely	(1/1)
Strident	(4/4)
Stringy	(5/7)
Strong	(5/7)
Subdued	(1/1)
Sweet	(4/4)
Thick	(2/2)
Thin	(12/14)
Timid	(1/1)
Tranquil	(1/1)
Tubby	(2/2)
Twee	(1/1)
Uninviting	(1/1)
Vibrant	(1/1)
Warm	(5/9)
Weak	(4/4)
Weighty	(1/1)
Woolly	(1/2)

Table 4.3 – Adjectives gathered with frequency of occurrence information

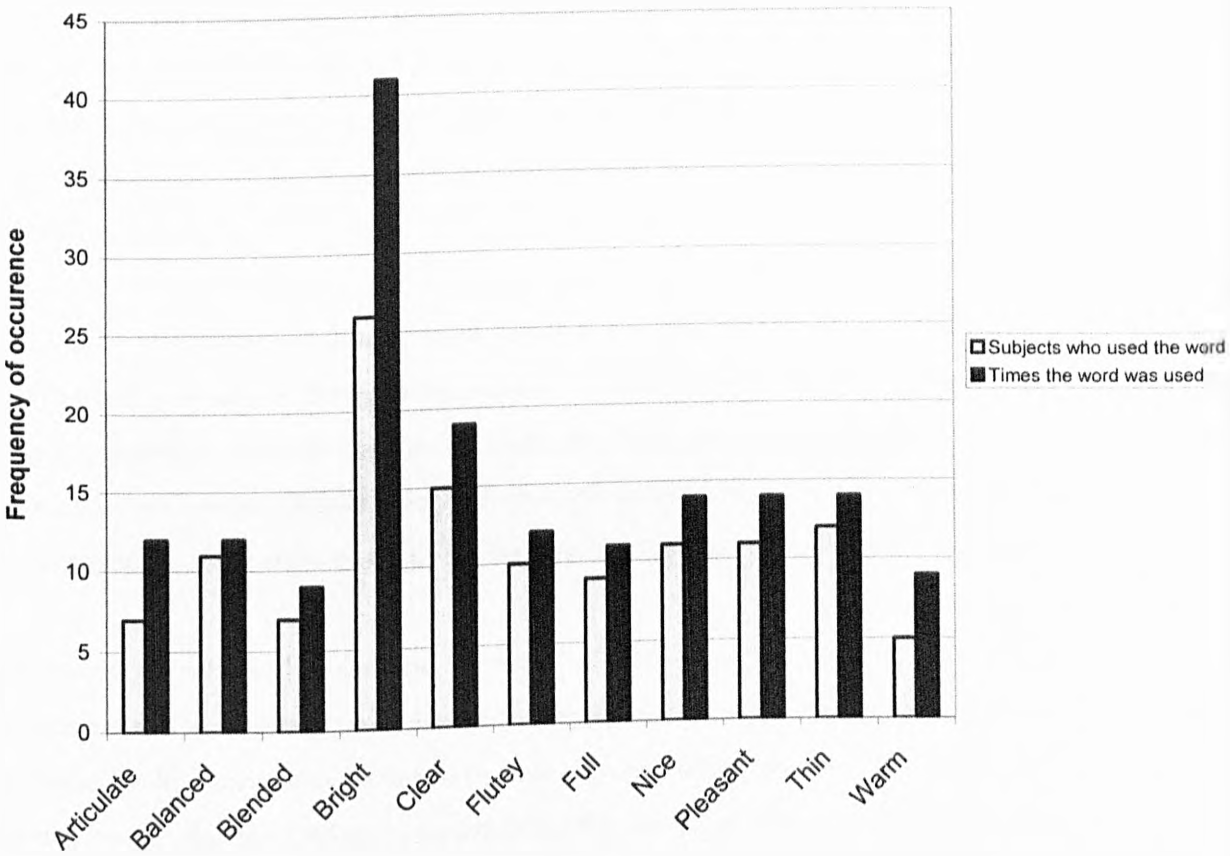


Figure 4.1 – Most common adjectives gathered by frequency of occurrence

A number of words require greater explanation. Dead and resonant might usually be expected to refer to the acoustic of the room rather than the sound of the organ, and answers where this was clearly the case have not been included in the figures above. Other words, such as "neo-Baroque", might be expected to refer to the casework, so only those listeners who were clearly describing the sound as "neo-Baroque" have been included in the results.

Some words, such as "reedy" or "stringy", refer to particular qualities of the pipe organ. As described in section 1.4.1, pipes of narrower scale are called strings in the organ, so a "stringy" ensemble might be more narrowly scaled. A "reedy" ensemble could either have a reed stop present, or other reed-like harmonics such as a Tierce stop not normally present in a principal ensemble. "Reedy" and "Stringy" were not mutually exclusive, and were used as synonyms by some people in the context of describing a principal ensemble. "Flutey", similarly, might suggest the presence of more flute-like tone in the ensemble. Words such as "chiffy" refer to qualities directly associated with pipe voicing.

Some words are clearly value judgements, such as "nice" or "pleasant", which do not describe the sound of the organ but the listener's opinion of it. Some listeners also described certain ensembles as "interesting". Such words are unclear in their meaning and probably have no direct connection with specific timbral qualities, being more abstract in nature.

It was also interesting to note that some words frequently used by acousticians that have well-defined meanings in those circles, such as "rough", occur infrequently here. Several people at a research conference, including Johan Sundberg, inquired of the author why such better defined and studied words had not been used for later stages of this experiment. The answer is that the listeners did not use such words at this stage.

The listeners were asked to describe the sound of the organs they were hearing. Many interpreted this in different ways, describing other features in addition to the sound of the organ itself. Some subjects talked of how each pipe might be scaled, and how the chorus was developed. Others commented on the suitability of the organ for the music being played, and the stops that might be in use. Any perceived unbalance in the sound that the listeners thought was due to a particular registration or voicing preference was also commented on by some.

Some listeners extended their comments on the organs by describing the acoustic each organ was in. Some subjects talked of how the sound projected into the room and some used a smaller variety of adjectives (e.g. "muddy") to describe the acoustic itself. Several combined this with other factors to guess at where each organ was placed. Where this was done, about half the guesses simply said "England" and the other half mostly guessed continental locations. A few subjects thought some of the organs were American. A few listeners made comments on or more precise estimates of the age and style of the organ in question.

Listeners were specifically asked if they recognised each organ from its picture, in case this affected their answers. If they did recognise the organ, they were asked to name it to verify their identification. In practice, because of the low recognition rate and imprecise nature of the data collected, it was not possible to make any meaningful analysis of this. None of the 31 American subjects thought they recognised any of the organs. Of the ten British subjects, three recognised all organs correctly, and three recognised just one correctly. Three also claimed to recognise two correctly, but one of those then misidentified the organ of the Jack Lyons as that in Kingston Parish Church (a more recent Frobenius instrument) and made no identification for the other instrument he claimed to recognise. The final British subject recognised none of the organs. The subject who lived in Greece but was originally English claimed not to recognise any of the organs, but did comment about the Jack Lyons instrument:

> "It sounded like a bright modern classical style instrument - a college perhaps? I went to look at the RCM site to see if it was the new room 90 organ, and I don't think it's the Lyon's concert hall."

No other subject thought they recognised any of the organs.

It is interesting to compare the general results in table 4.3 with the results gained from listeners in specific countries. Only the United Kingdom and United States were sufficiently well represented among listeners to make such analysis worthwhile.

Abrasive	(1/1)	Edgy	(1/1)	Pleasant	(2/3)
Aggressive	(1/1)	Elegant	(1/1)	Pure	(1/1)
Balanced	(1/2)	Fluffy	(1/1)	Reedy	(4/4)
Baroque (neo)	(1/1)	Flutey	(4/6)	Rich	(2/3)
Blended	(2/2)	Forced	(1/3)	Round	(1/1)
Body (has)	(1/1)	Full	(4/4)	Screechy	(1/1)
Breathy	(2/2)	Gentle	(1/1)	Sizzly	(1/1)
Bright	(6/11)	Hard	(1/1)	Small	(1/1)
Brittle	(1/1)	Harsh	(1/1)	Soft	(1/1)
Busy	(1/1)	Interesting	(3/3)	Solid	(2/2)
Buzzy	(1/1)	Lean	(1/1)	Sparkling	(1/2)
Character (has)	(1/2)	Light	(3/3)	Spiky	(1/1)
Chiffy	(2/3)	Lively	(1/1)	Steely	(1/1)
Clear	(4/4)	Mellow	(2/2)	Strident	(2/2)
Close	(1/2)	Metallic	(1/2)	Stringy	(1/1)
Cohesive	(1/2)	Muffled	(1/1)	Strong	(1/1)
Complex	(1/1)	Nice	(2/2)	Sweet	(1/1)
Creamy	(1/1)	Open	(1/1)	Thin	(2/2)
Dead	(1/1)	Piercing	(1/1)	Twee	(1/1)
Driving	(1/2)	Piquant	(1/1)	Warm	(1/3)
Dull	(2/2)	Plain	(1/1)	Weak	(1/1)

Table 4.4 – Adjectives gathered from U.K. tests subjects

Aggressive	(1/1)
Agreeable	(1/1)
Anaemic	(1/1)
Articulate	(5/8)
Authoritative	(1/1)
Balanced	(7/7)
Baroque (neo)	(2/3)
Bland	(1/1)
Blended	(4/6)
Body (has)	(1/1)
Breathy	(2/2)
Bright	(18/28)
Brilliant	(1/1)
Broad	(1/1)
Buzzy	(1/1)
Calming	(1/1)
Cerebral	(1/1)
Character (has)	(1/2)
Chiffy	(4/4)
Clean	(6/8)
Clear	(9/13)
Cohesive	(1/1)
Commanding	(2/4)
Competent	(1/1)
Crunchy	(1/1)
Cutting	(1/1)
Dead	(1/3)
Dignified	(1/1)
Distinct	(2/2)
Dynamic	(1/1)
Edgy	(3/3)
Elegant	(1/1)
Engaging	(1/1)
Exciting	(4/5)
Even	(1/1)
Firm	(2/2)
Flat	(1/1)
Flutey	(5/5)
Full	(5/7)
Full-bodied	(1/1)
Fuzzy	(1/1)
Gentle	(4/4)
Grand	(1/1)
Grating	(1/1)
Harsh	(2/2)
Heavy	(1/1)
Hollow	(2/2)
Homogenous	(1/1)
Integrated	(1/1)
Interesting	(4/4)
Jangly	(1/1)
Lazy	(1/2)
Light	(4/4)
Lively	(2/3)
Majestic	(2/3)
Mellow	(2/2)
Mild	(1/1)
Muddy	(1/1)
Nasal	(2/4)
Nice	(6/9)
Open	(2/2)
Overbearing	(2/2)
Piercing	(1/1)
Plain	(2/2)
Pleasant	(8/10)
Punchy	(1/1)
Pungent	(1/1)
Quiet	(1/1)
Reedy	(4/5)
Resonant	(2/2)
Restful	(1/1)
Rich	(5/5)
Robust	(2/2)
Rough	(1/1)
Round	(2/2)
Scratchy	(1/1)
Shimmery	(1/1)
Singing	(2/2)
Sluggish	(1/1)
Small	(1/1)
Smooth	(2/2)
Soft	(1/1)
Solid	(2/2)
Sparkling	(1/1)
Strident	(1/1)
Stringy	(4/6)
Strong	(4/6)
Subdued	(1/1)
Sweet	(1/1)
Thick	(2/2)
Thin	(10/12)
Timid	(1/1)
Tubby	(1/1)
Uninviting	(1/1)
Vibrant	(1/1)
Warm	(4/6)
Weak	(3/3)
Weighty	(1/1)
Woolly	(1/2)

Table 4.5 – Adjectives gathered from U.S. tests subjects

Of the common adjectives in figure 4.1, only one does not occur in both U.K. and U.S. subject groups. Only U.S. and Canadian listeners used "Articulate". "Clean" was the only word that occurred more than four times to be used exclusively by U.S. listeners. Many words occurred in descriptions from only one country, but these were mostly used by only one subject.

United Kingdom subjects used 63 different words, an average of 6.3 unique words per subject. United States subjects used 99 different words, an average of 3.19 unique words per subject. There could be a number of different explanations for this. U.K. subjects, although they were not given any different instruction, mostly took the test by direct personal request of the author, and therefore may have had more idea of what was desired based on their knowledge of the author's work. However, it is more likely that there is a core of common words, outside of which other words tend to be used by only one or two subjects. It would be expected that, as any subject group increased, the number of unique words per subject would decrease. Theoretically, therefore, the Canadian subjects, or any other small subset of subjects, should have a high unique word average. The Canadians did have an average of 6.33 unique adjectives per subject, but with only three subjects this is really too small to do any meaningful analysis on.

One possible reason for the use of certain words over others is the length of the word in question. Subjects might be more likely to use a shorter word than a longer word. Figure 4.2 shows the word length of all 129 words gathered plotted against their frequency of occurrence (excluding repeated use by the same person). Some of the points represent more than one word.

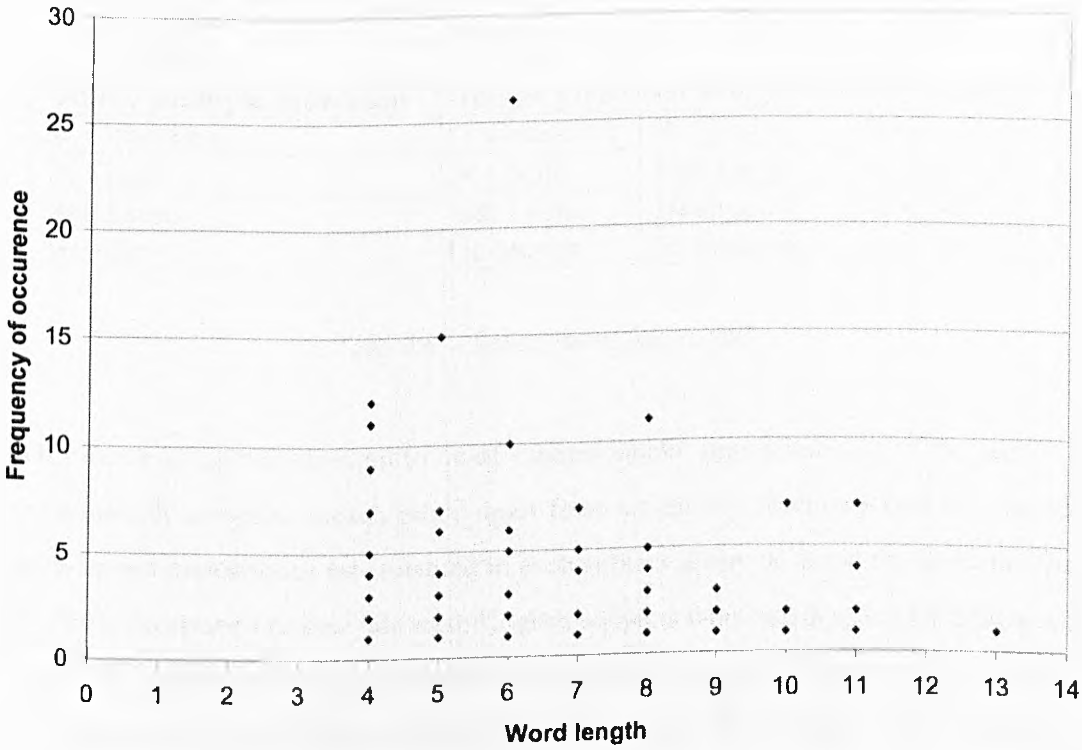


Figure 4.2 – Word length of adjectives gathered against frequency of occurence

Relevant values and other analyses can be found in appendix D. Although a tailing off of frequency of occurrence is evident with increased word length, this is probably consistent with wider patterns in English speech. As there are no factors evident that might have relevance to this research, further consideration of this aspect of semantics is inappropriate here.

4.5 The significance of displayed images on adjectives used

As introduced in section 4.3.2, the test subjects were divided into three groups, each of which was presented with the samples in the same order but the images in different orders. The images used were of the four organs in question, and can be found in appendix E alongside the descriptions of the organs used. The research covered in this section was described in Disley (2003). The size of each subject group and the pictures presented to them are detailed in table 4.6.

	Group 1	**Group 2**	**Group 3**
Number of subjects	17	18	17
U.K. Subjects	4	5	2
U.S. Subjects	12	10	13
Audio example presented	**Image presented alongside audio example**		
St. Columba's	St. Columba's	St. Chad's	Heslington
St. Chad's	St. Chad's	Jack Lyons	St. Columba's
Jack Lyons	Jack Lyons	Heslington	St. Chad's
Heslington	Heslington	St. Columba's	Jack Lyon's

Table 4.6 – Subject subset group details

Each subject group was selected to be of experimentally significant size. Volunteers were randomly assigned to each group apart from an attempt to ensure that the overall ratio between nationalities was retained in each subject group to avoid the potential for bias. Two exceptions to this rule were English subjects who had detailed knowledge of some of the organs in the test, and thus would be likely to spot discrepancies between the images and samples if these were of different organs. Those subjects were placed in the first group. A small number of subjects were unable to complete the test, which is why the groups and geographical subsets thereof are not quite as similar in number as they should have been. In two of the failure cases, the problem was due to the subject's computer being unable to play the sound files. In the other cases, for various reasons the volunteer did not have time to complete the test. In percentage terms, U.S. volunteers were more likely to successfully complete the test (86% completion) than U.K. volunteers (77%) and other volunteers (83%) although the latter subject group was too small (6) for meaningful comparison. The lack of problems due to the test interface suggests that the alterations in experimental procedure have been successful.

As mentioned in section 4.3.2, a fourth subject group with the remaining audio and visual combination possibilities was undesirable. Such a group would have made the unlikely combination of the sound of the Jack Lyons organ with the picture of the St. Columba's organ, and the sound of the St. Columba's organ with the picture of the Jack Lyons organ. The former would be unlikely, although possible in the case of a new organ in an old case, but the latter would have been completely improbable. Had sufficient subjects volunteered, it might have been interesting to see how they reacted to this perceptual conflict, but with the number of volunteers as they were, it was preferable to ensure that the other three subject groups had adequate numbers.

As the results of this section are not numeric in nature, subject responses for each organ will be summarised before overall trends are analysed. Full responses, divided by test and organ, can be found in appendix C. Numbers in brackets after comments in this section refer to the numbering of subject responses in that appendix. The same number can refer to answers from different subjects in different tests.

4.5.1 Subject responses to the recording of St. Columba's

When the recording of St. Columba's was presented with the correct picture, there were both common trends and disagreement. Six (of seventeen) subjects commented on the dead acoustic (2, 3, 9, 12, 13, 15). Three suggested that the extract was played on 8' and 4' principals (8, 11, 15). Two correctly identified a 2' component (7 & 13), but one thought this was overly prominent. Four commented on its appropriateness for congregational or parish church use.

When combined with a picture of St. Chad's, again people thought the sound was average. Several people commented on a prominent 2' component (4 & 9), or a lack of mid-range sound in comparison to the bass and treble (15). Overall there was little difference from the response to St. Columba's with its own picture. Similar comments were also received from the combination of the sound of St. Columba's with the picture of Heslington. Overall, there was little difference between the three subject subsets in their comments, although individual subjects within each subset disagreed particularly on whether the organ was good, bad or indifferent.

4.5.2 Subject responses to the recording of St. Chad's

When subjects were presented with both the picture and recording of St. Chad's, the responses were far more favourable towards the organ, although one or two had reservations. Words such as "exciting" and "well-balanced" were used. Eight subjects (4, 5, 10, 11, 12, 13, 15 & 17) commented on the more reverberant acoustic and how it helped the organ tone. Four subjects (1, 7, 14 & 15) commented on a Tierce or Twelfth/Nazard component (or a mixture that contained these, such as a Cornet or Sesquialtera). Six subjects (5, 10, 11, 14, 15 & 17) described the ensemble as either reedy or probably including a chorus reed. There was disagreement as to the organ's date, with 18th, 19th and early 20th centuries all being suggested. No subjects suggested a later date, but one thought the sound was from an earlier organ than the case.

When the sound of St. Chad's was combined with the picture of the Jack Lyons instrument, although comments on the reedy plenum and good acoustics were similar, subjects were less complimentary about the organ's balance and blend. Two subjects (6 & 10) described it as a neo-Baroque or Baroque revival sound, and another (4) dated the sound as contemporary. None thought the sound inappropriate for the organ pictured.

When the picture of St. Columba's was displayed with the sound of St. Chad's, two subjects (1 and 13) noticed a mismatch between the picture and the sound. Many commented on its brightness, and the negative comments were mostly directed towards the acoustic.

4.5.3 Subject responses to the recording of Jack Lyons

When both picture and recording were of the Jack Lyons, subject responses were mixed. Many commented that they were disappointed, and suspected the instrument had greater resources. When the picture was of Heslington, although the comments on tonal make-up were similar, none thought that the sound was inappropriate for that organ. One subject (8) described the sound as "not neo-Baroque", although others commented on the apparent lack of nicking and open foot voicing (7). One subject (6) even described the sound as "quite romantic".

When the sound of the Jack Lyons was matched with the image of St. Chad's, many again commented on its flutey sound. One subject (8) described the sound as baroque, and two (3 & 9) made comments on its suitability for congregational use. One subject (13) thought there was a mismatch between the case and the instrument now in it, although his comment suggests that the perceived mismatch may have been visually stimulated.

4.5.4 Subject responses to the recording of Heslington

Subject responses to Heslington when its picture and recording were combined were mostly favourable, typically commenting that it had an unforced, pleasant sound, although no subjects were overly enthusiastic. One subject (6) commented that the instrument was probably recycled, although whether it was case or sound that prompted this is unclear.

When the sound of Heslington was combined with the picture of St. Columba's, some subjects again were broadly favourable, but four subjects (2, 4, 6 & 11) commented on a mismatch between the picture and the sound. Several also disliked the sound, although making no reference to the casework.

When the sound of Heslington was played alongside pictures of the Jack Lyons instrument, the comments were brief, but all subjects appeared positive in their responses. None thought the case inappropriate for the sound played. One (13) thought the instrument "clearly American", others identified it as "Massachusetts" (2), "Kingston Parish Church" (11) and "North German in nature [of scaling]" (4).

4.5.5 Analysis of subject responses

Most of the subject answers were descriptive and did not relate the sound heard to the organ case shown. However, in a few cases, particular combinations of picture and sound produced interesting responses, and it is these that give greater insight into the effect of the displayed image upon subjects' perceptions of sound.

The sound of St. Chad's was described as neo-Baroque when presented with a picture of the Jack Lyons, and as much older when accompanied by other pictures. The Jack Lyons sound, when played alongside images of other organs was both described as baroque by one subject and not neo-Baroque by another subject. However, not too much can be directly derived from this result, as the Jack Lyons ensemble chosen was not perhaps the most obviously neo-Baroque of its potential registrations. In all cases, the results and comments only refer to the particular registrations chosen here, and any conclusions about individual organs cannot be extrapolated to all its available registrations.

In a number of cases where there was a perceived mismatch between the casework and sound identified by some subjects, other subjects who did not mention a mismatch still reacted negatively to the instrument. The same ensembles gained more positive reactions when accompanied by other pictures. This suggests that a conflict between the sound an instrument makes and the sound it might be expected to make based on its case (and possibly surroundings) may generate a more critical or negative response in listeners, even if the mismatch is not consciously identified.

It is often difficult to categorise subject responses as good or bad, but most of the responses were either mostly positive or negative towards the instrument in question. The two most consistently positive reactions were to the recording of St. Chad's accompanied by its own picture, and to the recording of Heslington when accompanied by a picture of the Jack Lyons instrument.

Several subjects made comments on the scaling of particular ensembles, but these did not always agree. When subjects noted their playing and listening experiences, sixteen mentioned some degree of experience at organ building. It is clear from the comments that as well as disagreement over how a particular ensemble is scaled and balanced, there was disagreement over whether that ensemble was scaled well. This should not be surprising, given the amount of disagreement generally as to what was good or bad.

In a few cases, subject answers appear to indicate a mistake on the part of the subject in identifying the size of the case. Subjects who guessed the location were also largely wrong, apart from those who made the logical deduction that all the organs might be local to the author and those who knew the organs in question.

4.5.6 The effect of memory and elapsed time on organ identification

One final consideration when deciding whether to display images was whether a previous combination of an organ's sound and picture, as presented in this part of the research, would be remembered and potentially bias future results. To determine the effect of subjects' memories of connections between audio samples and the images presented alongside them, a further test was devised. Such memory in the context of these tests would not necessarily be identical to subjects' memories of real-world organs, due to the limited samples presented of both the organ's appearance and sound.

Once subjects had taken the previous test, they were presented with the four samples and pictures of the four organs, and asked to match organs to samples. Subjects were free to match several samples with the same picture, but not to choose more than one picture for the same sample. Some subjects took this test immediately after taking the other test, and others took it at intervals of up 57 days.

Of 46 subjects who took part in the adjective gathering portion of this research, 29 went on to complete this memory experiment. Those who did not take part were mostly those for whom significant time had elapsed between the two experiments. Eighteen subjects took the test immediately after the adjective gathering experiment, at intervals of between 1 and 48 minutes between submission of the two parts. Eleven subjects took the test at greater intervals, between 12 and 1737 hours after first taking the test.

Due to the way in which subjects were allocated to different groups in the adjective gathering experiment, most of the subjects who took this test immediately were members of group 3. The subject distribution can be found in table 4.7.

Number of participants who took the test:	Subject group		
	Group 1	Group 2	Group 3
Immediately	2	5	10
Delayed	7	4	1

Table 4.7 – Previous grouping of participating test subjects

One subject from group 3 gave the same answer for all four recordings. Subsequent correspondence demonstrated that he had not intended to do so, so his answers were disregarded in the overall analysis. Subjects' answers were compared against two sets of potential answers: the correct combination of images and sound samples, and the combination of images and sound samples that the subjects had been given. In the case of the first subject group, these were the same. The number of correct answers for each subject group is shown in table 4.8.

Number of correct answers		Subject groups		
		Group 1	Group 2	Group 3
For the actual combination of image and audio	4	6	0	0
	3	0	0	0
	2	1	1	0
	1	1	2	5
	0	1	6	6
For the combination of image and audio presented to that group	4	6	6	4
	3	0	0	1
	2	1	2	3
	1	1	1	2
	0	1	0	1

Table 4.8 – Number of correct identifications divided by subject group

Groups 1 and 2 exhibit similar patterns of memory, with the majority of subjects remembering the combinations of audio and image presented to them. Group 3 seems somewhat worse in this regard, which is unexpected as all but one of group 3 took the test immediately, and thus might be expected to exceed the other groups in remembering the combinations presented to them. The members of group 3 were no more likely than those of group 2 to have identified potential mismatches in the previous test, and thus deliberately choose combinations other than those originally presented to them. It is possible that the immediate presentation of the second test suggested the likelihood that organs in the previous test had not been correctly matched with their sounds, but this would also have been the case for some in the other subject groups, and is not borne out in results.

Overall, the sooner this test was taken after the first test, the more likely subjects were to identify the combinations of audio and visual stimuli previously presented to them. Those who did not identify all their previously presented combinations made no statistically significant move towards the real-life combinations. It would be expected that a random choice would produce on average one correct answer in each combination category in table 4.8.

The effect of time on the number of identifications of image and audio combinations previously presented to subjects is shown in figures 4.3 and 4.4. Figure 4.3 shows those who took the test within fifteen minutes of the previous test, and figure 4.4 shows those who took the test after a delay of at least 45 minutes. Some data points represent more than one answer, and full results can be found in appendix F.

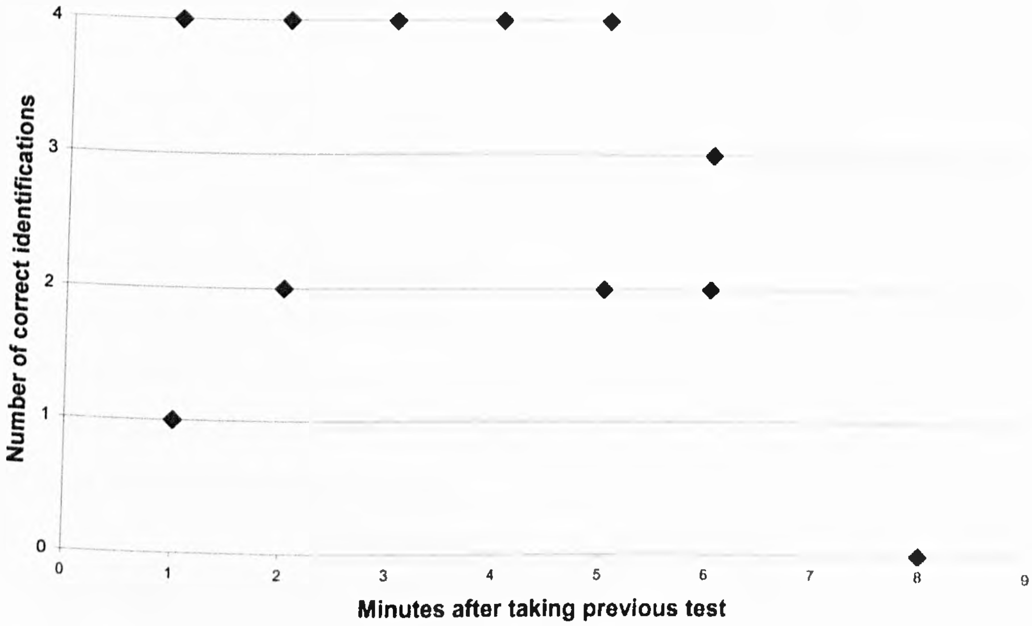


Figure 4.3 – Number of correct identifications over short-term time

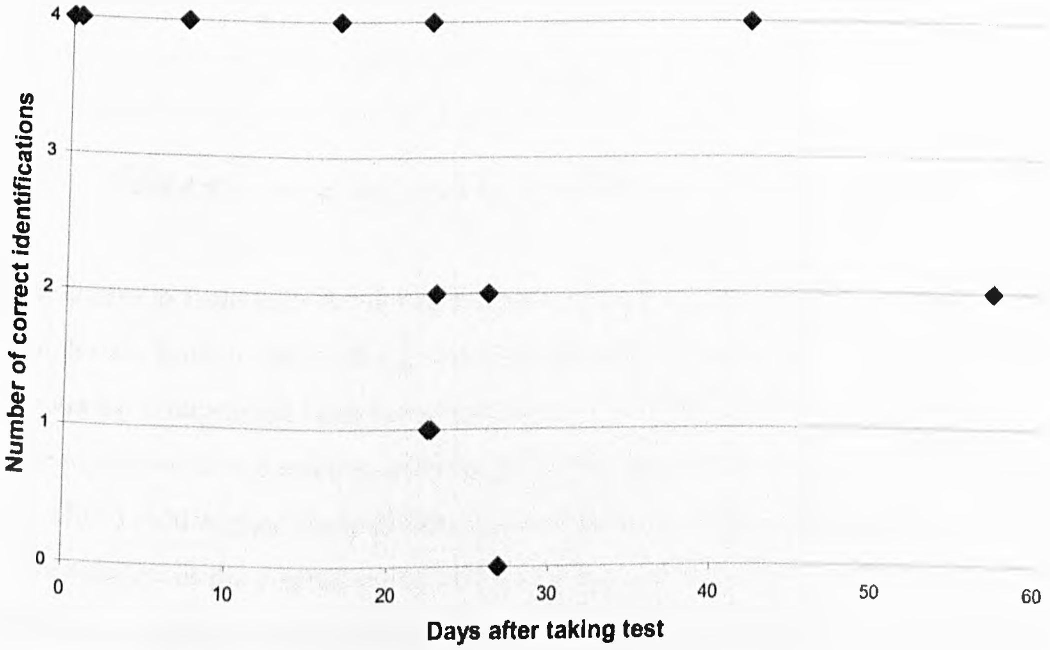


Figure 4.4 – Number of correct identifications over long-term time

It is not clear from figures 4.3 and 4.4 exactly what the effect of a delay between taking these two tests is. Figure 4.3 would seem to imply a falling off over time, but as figure 4.4 illustrates, some subjects were able to correctly identify all four combinations presented to them after a period of time had passed. Those who did so after more than a day were all from the first subject group, and two of the four knew some of the organs in question. However the other two subjects, including the correct identification with the longest delay of nearly 42 days, did not know the organs in question. It may be more illuminating to look at the average retained recognition ratings to see if there are common trends not evident from the scattergrams.

Number of combinations correctly identified	Average time passed (in hours)	Standard deviation	Number of subjects
4	131.8	282.0	16
3	0.1	n/a	1
2	423.2	546.5	6
1	359.2	311.0	3
0	323.0	456.6	2

Table 4.9 – Average time passed for all possible numbers of correct identifications

What is certain from table 4.9 is that subjects incorrectly identifying two or more examples are likely to have left a greater gap between taking the two tests. It cannot however be extrapolated from this result that an increasing amount of time passing between such tests will result in subjects being less likely to successfully complete the test. This would require study of subject numbers for each combination, and with only twelve subjects in the delayed group and no clear trends in their data, it would be dangerous to assume a roll-off in accuracy over time based on this data. It is clear that some subjects have the ability to remember the combinations of audio and image presented to them over considerable periods of time.

4.5.7 Conclusions on the significance of displayed images

In conclusion, it appears that the picture displayed does have an effect upon the words used to describe the sound of a particular organ. A mismatch between the organ displayed and that heard, whether consciously noted or not, can produce a more negative and critical reaction to the ensemble. Some subjects can also remember the use of pictures and their combination with a particular ensemble for considerable periods of time. Thus the use of a particular picture might prejudice its accompanying ensemble if it had been displayed alongside a different registration for that organ or the sound of an altogether different organ. Pictures, whether of the organ in question or not, therefore should not be used in future experiments to avoid the possibility of biasing the result.

As this part of the experiment was based upon analysis of verbal descriptions, the precise effect of an image cannot be quantified. Possible future work to provide such quantification could involve the comparison of two or more subject groups' ratings on various scales. From this work it can be surmised that some of those scales would probably produce different results if a different picture was presented.

4.6 Refining and classification of adjectives

Although the adjectives gathered thus far had a number of clear candidates for further study based on frequency of use, some of those common words would have less practical use than others. Words such as "nice" are clearly based not on a particular common timbral quality of the sound but on a higher perceptual layer of personal preference. The individual comments in the adjective gathering test have shown that such personal taste varies dramatically. To choose useful words for further study, some means of identifying this usefulness needs to be devised.

Previous studies have categorised adjectives gathered or studied arbitrarily, presumably based on the authors' opined etymology. At a conference, the author of this research did likewise with the adjectives gathered in section 4.4. Several other researchers suggested that such arbitrary methods might merely reflect the author's personal view. In light of this advice, the author opted to present the most common adjectives to a subject panel for classification, a method not previously undertaken in the field of timbral semantics. Subjects were gathered from a similar field, which also enabled the author to ask whether each word was useful in the context of describing the sound of a pipe organ.

Rioux (2001) and Creasey (1998) summarise much of the existing categorisation of adjectives in previous studies. For example, in Rioux (2001, p33) are details of research by Solomon (1958) in which he placed pairs of words (related to words derived from sonar) in the following categories:

> Magnitude, aesthetic-evaluative, clarity, security, relaxation, familiarity and mood, as well as an unidentified option.

Rioux went on (2000) to place various descriptors relating to the speech of individual pipes in two sets of inter-linked categories:

- Speech or transient, steady state, general impression, and proportion between various tonal qualities
- Onomatopoeia, antonym, and physical parameters.

In Rioux and Västfäll (2001a), the process of categorisation is not entirely clear, but appears to have been conducted by the authors adding categories whenever they felt an adjective gathered did not fit in an existing category. Subjects were not involved, and some words that did not fit in any category were discarded for that reason alone. The complete list can be found on pages 108-109 of Rioux and Västfäll (2001a), and is similar to those listed above with the additional of a "metaphor" category.

Each experiment has its own set of categories devised by the author(s) to meet its particular need. In the case of the research described in this thesis, it was important not to confuse the subjects either with difficult to define categories or a multiplicity of category dimensions. As five sense-based categories had already been used by the author, these were used again in this instance to compare the author's own categorisation with that of the test subjects.

4.6.1 Experimental procedure for adjective classification

Sixty-six subjects were recruited from the Internet mailing list piporg-l and the author's personal contacts. Some had previously taken part in the author's research, and others had not. Subject demographics are presented in tables 4.10 and 4.11.

Language spoken	Number of subjects
English (British)	11
English (Canadian)	1
English (American)	54

Table 4.10 – Language spoken by participants in the adjective classification experiment

Playing experience	Number of subjects
No playing experience	1
Limited playing experience	9
Good organist	25
Professional organist	31
Other relevant interests mentioned by subjects (not necessarily mutually exclusive)	
Professional organ builder	4
Amateur organ builder or limited experience of organ building	13
Pipe organ researcher	5
Electronic organ technician	1
Pipe organ music broadcaster	1

Table 4.11 – Relevant experience of participants in the adjective classification experiment

As the results of the previous free choice method of playing ability had proved difficult to interpret, a fixed choice method was reverted to with revised categories listed in table 4.11. It should be noted that of the eleven subjects who described themselves as speaking English (British), two came from Canada, two from Australia, and one each from South Africa, Greece and Spain. The latter two were known to the author to be English ex-patriots, but it is likely that some of the others chose that option as the nearest to their language. English (Canadian) was not an option, but one subject described his language thus in the "Other" option box. It is also possible that some of the English (American) subjects were from Canada, but none had the geographic domain identifier ".ca" on their email addresses that permitted identification of the two Canadians above.

Despite these regional complexities, what is clear is that all subjects speak English as a first language, and the majority are from the North American continent. There are insufficient definite UK English speakers to permit any meaningful comparison between their answers and those of the US English speakers at this stage.

Subjects were presented with all fifty-nine adjectives that were used more than once in the first section of this adjective gathering experiment. The intention was not to include those words only used by one subject, but four such words (Close, Lazy, Shrill and Woolly) slipped through this check. Additionally, four words used by more than one subject (Full, Interesting, Small and Romantic) were not included by error. "Full" was perhaps the most serious omission, the mistake perhaps due to it being sufficiently popular that its inclusion was assumed rather than checked. "Interesting" was not originally counted as a timbral adjective, but on later consideration was included in the tables above. "Romantic" and "Small" were subject to a simple counting error when the original subject responses were analysed that went unidentified until after the classification experiment.

Subjects were given the following instructions:

> The words on this page were used by listeners to describe the sound of a number of different pipe organs (yes, even the strange words!). This experiment is two-fold: to discover which category you would put each word in, and to discover whether you think that particular word is useful when describing the sound of a pipe organ. The categories are:
>
> - Taste or Smell - if you think the words originate from the sensation of taste or smell (e.g. "sour", "scented" or "ripe")
> - Emotions - words such as "intense" or "melancholy"
> - Vision - words such as "luminous"
> - Sound - words that are gathered from acoustics or hearing
> - Touch - words that originate with tactile sensation
> - Other - if you don't think the word fits in one of those categories. Please use the text box to describe an alternative category (or indeed a second category) if you have one in mind, but don't feel that you have to use the text box.
>
> Please go with your instinctive first answer, and don't think about anything too much. When considering whether you a word is useful for describing the sound of a pipe organ, please remember that you might be describing both good and bad ones.

4.6.2 Results of adjective classification

The results are presented in table 4.12, which has been split across multiple pages to permit better comprehension and legibility. The order of adjectives is the same as that presented to subjects, and is in the order of frequency of occurrence in the adjective gathering experiment. Numbers marked in bold are the categories that the author originally placed the words in. Adjectives marked in bold are among the list of 85 presented to subjects by Rioux and Västfäll (2001b, p125), and many of the adjectives not present on their list had synonyms that were present.

Adjective	Sound	Vision	Emotion	Taste/ Smell	Touch	Other	Useful
Bright	38	**31**	0	0	0	0	64
Clear	38	**29**	**1**	0	1	0	60
Pleasant	15	3	**48**	4	2	3	42
Thin	52	**4**	1	2	9	1	59
Flutey	**66**	1	0	0	0	0	62
Warm	28	3	22	4	**16**	2	59
Balanced	41	5	7	1	4	**15**	56
Blended	47	3	4	**8**	2	8	50
Articulate	**52**	2	0	0	11	5	50
Reedy	**64**	2	0	0	1	0	61
Nice	8	4	**51**	3	2	7	23
Baroque	46	5	1	0	2	**15**	60
Chiffy	**64**	1	1	0	2	0	63
Clean	29	14	4	13	**8**	4	48
Rich	42	4	11	**14**	1	2	55
Smooth	29	2	5	5	**28**	3	52
Stringy	**57**	2	0	2	6	2	58
Exciting	8	3	**55**	2	2	5	37
Mellow	35	2	27	**4**	1	3	57
Strong	18	5	32	8	4	**7**	36
Breathy	**59**	0	4	1	1	2	54
Gentle	18	2	34	2	**43**	2	45
Harsh	**38**	2	23	3	5	2	55
Lively	15	3	**40**	2	4	9	31
Nasal	**65**	0	0	1	1	0	55
Resonant	**61**	0	2	0	0	4	53
Strident	**47**	0	14	0	0	5	55
Characterful	13	1	**39**	2	1	14	55
Cohesive	36	4	9	1	5	**15**	47
Commanding	24	2	**36**	1	1	7	42

Table 4.12 – Adjectives classified by category with usefulness rated (continues on next page)

Adjective	Sound	Vision	Emotion	Taste/ Smell	Touch	Other	Useful
Dead	33	1	**26**	2	4	5	44
Distinct	39	**10**	6	6	6	8	41
Dull	41	**6**	19	2	6	0	55
Edgy	30	0	29	1	**8**	2	39
Light	25	27	4	2	**12**	3	47
Open	32	**8**	17	0	0	9	33
Piercing	**58**	2	4	1	4	1	60
Soft	46	2	7	1	**17**	0	54
Solid	22	4	10	1	**29**	6	42
Sweet	26	2	20	**21**	2	3	44
Weak	30	2	21	4	**10**	7	45
Brilliant	44	**16**	4	0	0	2	63
Buzzy	**63**	0	0	0	2	1	57
Close	11	7	16	0	**17**	16	15
Elegant	12	**12**	33	3	6	12	39
Firm	11	1	9	1	**45**	3	32
Heavy	23	0	7	4	**31**	5	48
Hollow	54	**1**	3	0	5	4	54
Lazy	4	0	**45**	1	3	13	6
Majestic	30	**3**	32	0	1	4	56
Muddy	44	**14**	2	2	4	4	56
Plain	18	**16**	12	8	1	15	27
Robust	36	2	15	**9**	4	6	49
Round	34	**11**	5	0	7	10	45
Shrill	**64**	0	2	0	0	0	61
Singing	**61**	0	4	0	0	1	50
Sparkling	33	**24**	7	4	0	1	58
Thick	37	2	6	1	**20**	4	50
Woolly	28	2	2	3	**24**	8	36

Table 4.12 – Adjectives classified by category with usefulness rated (continued from previous page)

Table 4.12 includes second categories where subjects used the "other" box to indicate this, so results from the central six columns may add up to more than the number of subjects.

4.6.3 Alternative adjective categories suggested by subjects

The alternative categories that subjects suggested were interesting for a number of reasons. Some suggested a misunderstanding of what was being asked of them, and others clearly demonstrate that the sense based categories chosen by the author do not adequately describe the origins of all the adjectives gathered. In the following consideration of adjective categories suggested, words and numerals in brackets refer to the adjectives thus classified and the number of people to do so. Where no number is given, only one subject gave that response for that adjective.

A number of responses classified certain adjectives as pertaining to the physical world. While such qualities may be interpreted via one of the sense categories given, subjects clearly felt that the attribute was primarily outside of human perception. Alternative categories suggested for this area were:

- Size (thin)
- Weight (thin, balanced (2))
- Temperature (warm)
- Kinaesthetic (balanced, strong, lively, round, robust, lazy, weak)
- Physical [properties] (strong (2), breathy, light, round (2), hollow, thick, weak, solid)
- Physical percept (open, close)
- Motion / movement (lively, strident)
- Referring to space or area (open, round, close, thick)
- Environment (close)
- Condition (dead)

Several other adjectives were classified as not pertaining to any particular sense but being products of a higher level of human intellect or reasoning. While some of these categories here and elsewhere may often seem synonymical, all variants are presented to preserve subjects' original nuances of meaning. Suggested categories were:

- Intellect (balanced (2), Baroque, characterful, round, elegant, lazy, cohesive)
- Analytical (balanced, Baroque, cohesive, elegant)

- Intuition (balanced, blended, cohesive)
- Logical (balanced, cohesive)
- Cognition (Baroque)
- Proprioception [sic] (balanced)
- Attitude (nice)
- Opinion (Baroque)

Several suggested categories were more artistic or historical in nature. These words included:

- Style [type of] (Baroque (2), elegant, characterful)
- Architecture (Baroque, close)
- Historical [period] (Baroque (5))
- Overly decorated (Baroque)
- Art (Baroque)
- Proportion (thin)

Some subjects gave more precise categories for words that could be said to lie within some of the sense categories provided. Such words were:

- Cookery (blended)
- Chemistry (blended, cohesive)
- Whisky (blended)
- Wind (breathy)

Some subjects were unable to think of a suitable adjective category. In one exceptional case, the subject rated fifteen words as "neutral", presumably meaning that they did not clearly fall into any particular category, whether provided or not. These words were:

> Pleasant, blended, nice, smooth, exciting, mellow, strong, gentle, characterful, cohesive, commanding, distinct, solid, plain, close.

Another subject used the slight oxymoron of "descriptive adjective" as a category for the following eight words:

> Nice, lively, open, round, plain, woolly, close, lazy.

A small number of subjects described words as "vague", or put "none", "Can't think of a category" or just "?" in the alternative suggestion box. Adjectives thus described were:

> Characterful (2), distinct, muddy, plain (2), woolly, close (3), lazy.

Two words were also criticised by a few subjects. One described "firm" as "not a good sound term". Two responded to "characterful" with "bad word" and "this is not a word". Certainly it is not to be found in the Little Oxford Dictionary, but 56 subjects categorised it, and 55 subjects thought it a useful term. In retrospect, perhaps an alternative could have been found to "characterful" as an attempt to avoid the inelegant pseudo-adjective "having or possessing character".

While a lack of categorisation may initially be disappointing, it demonstrates the difficulty some subjects had in categorising the less obvious words. It is interesting to compare the words here with the usefulness ratings in table 4.11. Of the words that prompted two or more subjects to give a "don't know" type answer, few have high usefulness scores:

> Lazy (6), close (15), nice (23), plain (27), woolly (36), distinct (41), and characterful (55)

Of these, "lazy", "close" and "nice" are significantly below the mean usefulness score (48.31, with answers below 24.72 being statistically significant at a 5% level). However, a lack of a convenient alternative category is not necessarily indicative of a lack of use or usefulness. One subject commented about the adjective "muddy": "I found this difficult to categorise, though have used it to describe organ sound".

Instead of providing alternative categories in the "other" box, several subjects suggested synonyms and definitions. These alternative words were:

- not harmful (pleasant)
- representing comfort (warm)
- all things in proportion (balanced)
- orderly, well-behaved (clean)
- abrasive (harsh)
- invigorating (lively)
- loud, unrefined (strident)
- captivating of attention (commanding)
- unique, all things readily perceived (distinct)
- all things gentle, airy (light)
- without obfuscation, broad (open)
- firm, reliable (solid)
- gentle, appealing (sweet)
- cohesive (close)

One subject who contributed most of the above used the category "other" 22 times out of 59 words. This was exceptional compared to the rest of the subject group, who used it on average 5 times each.

Many subjects, including some organ builders, used the "other" box to relate the adjectives in question specifically to the pipe organ, although this was not asked of them. Some words were thought to relate primarily to playing or performance technique, rather than the organ being played:

> Articulate, clean (2), smooth (2), exciting (2), lively, strident, characterful, cohesive, commanding, dull, brilliant, majestic, elegant (2), lazy.

Some commented specifically that the following adjectives would refer to particular choices of registration by the player rather than technique:

> Rich, commanding, muddy and woolly.

Some subjects thought that certain words would not be applicable to full ensembles, but only to individual ranks:

Mellow, strong, and hollow

A small number of subjects were more specific in their description of which type of pipes a word might refer to:

- Foundations: strong, cohesive, solid
- Flues: majestic
- Flutes: gentle, distinct, open, soft
- Reeds: lively, strident, buzzy, majestic, piercing
- Mutations: lively
- Cromorne: characterful
- Trompette: commanding
- Dulciana: sweet
- Voix [sic]: resonant

Some subjects thought particular words would refer to the voicing and tonal finishing of the organ. Again, although some people may consider voicing and tonal finishing synonyms, they have been left separate here to preserve any nuances of meaning intended by the subjects.

- Voicing: stringy, mellow, elegant (the latter referring to unforced voicing)
- Tonal finishing: resonant, strident, cohesive, dead, dull, solid, brilliant, heavy, majestic, muddy, hollow, lazy
- Pipe speech: articulate, lazy

One subject commented that "elegant" would be more likely to refer to the case than its contents. Some also suggested that the acoustic of the room was more likely to be described by these adjectives:

Resonant (3), strident, dead (3), distinct, muddy, hollow.

Finally, it was good to note that some subjects kept a sense of humour throughout the test. One commented that he would only use "woolly" in conjunction with "wild", another suggested "student" as a category for "lazy", and another suggested relegating "nice" to only be used for biscuits and France.

4.6.4 Additional adjectives suggested by subjects

Seventeen subjects suggested a total of 97 additional adjectives not included in the list given to them to categorise. Some were described as being appropriate in a particular context, such as that of the theatre organ. The words follow, with numbers in brackets where more than one subject suggested the same word:

> Appropriate, barking, biting, blaring (2), blatant, blatty, blue, booming, brassy (2), buttery, cheerful, coarse, colourful, complex, congested, contemporary, crude, dark, deep, discordant, ear-splitting, eerie, enfolding, English, ethereal, euphonious, exquisite, fiery, fifthy, fluffy, French, funky, German, glittery, glorious (2), gorgeous, grating, hard, harmonious, harsh, historic, hooty, humorous, icy, imitative, insipid, Italian, jarring, jazzy, leaden, lovely (2), lush (4), lyrical, melodious, moody, moving, musical, mysterious, murky, orchestral, penetrating, persuasive, piquant, presence, profound, raspy, raucous, regal, rolling, Romantic (3), sacred, sagging, screechy (2), shimmering, silvery, slow, snappy, spicky, spiritual, subdued, substantial, subtle, syrupy, tasteful, thrilling, thunderous (3), tinny, torchy, treacly, velvety, vibrant, visceral, well-ordered, whiny, wistful, woody (2), zippy.

Some of the words listed above are unfamiliar to the author, but are included here nonetheless. These words fill out those previously gathered. Some adjectives clearly refer to schools of organ building. Other words fit comfortably in categories both given in the test and suggested by participants. One subject commented "I could go on and on" – clearly there is no shortage of such words, and the large number gathered in this manner from only seventeen subjects concurs with Moravec and Štěpánek (2003) who obtained 1964 words from 120 subjects. Romantic, one of two words omitted from the adjective categorisation experiment by mistake, was suggested three times, confirming its potential as a useful adjective in some contexts. One subject suggested only that word, as an opposite for "Baroque". It is important to emphasise that these words were not gained in response to auditory stimuli, unlike the initial gathering of adjectives, and thus the presence in that list of a certain word does not necessarily mean that it would be used by the person who suggested it. As these words were not gathered in a formal manner, they are included here for interest only and will not be subject to further analysis.

Three subjects also commented on some of the words in the first part of the test, giving their own personal definitions thereof. While interesting, a common definition of a word does not equate to a common understanding, as has previously been demonstrated. For this reason, combined with the small number of words thus defined, none of which were the commonest adjectives gathered (as presented in figure 4.1), these definitions are not considered here.

4.6.5 Analysis of adjective classification

As the order of adjectives presented in the categorisation test was that of most to least commonly used, it is interesting to look at the usefulness scores against test position to determine if usefulness has a relationship with frequency of actual usage.

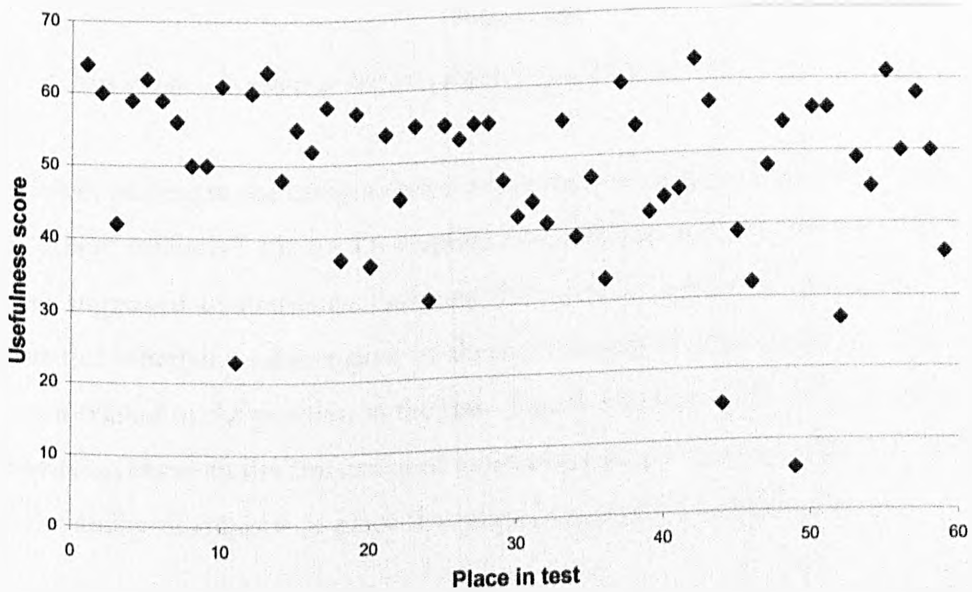


Figure 4.5 – Usefulness score against place in test

Some of the later words in the test, used by only two people in the adjective gathering experiment, still get usefulness scores comparable with earlier words in the test. What is visible, as the place in test measure increases, is not so much a smooth roll-off as an increase in the standard deviation of usefulness scores. Words occurring earlier in the test are more likely to be judged useful by a significant majority of test subjects. Words occurring later take on a more random pattern of usefulness scores. This means that it is hard to predict whether, for example, the alternative adjectives suggested by more than one subject (but no more than four) in section 4.6.4 would be considered useful or not.

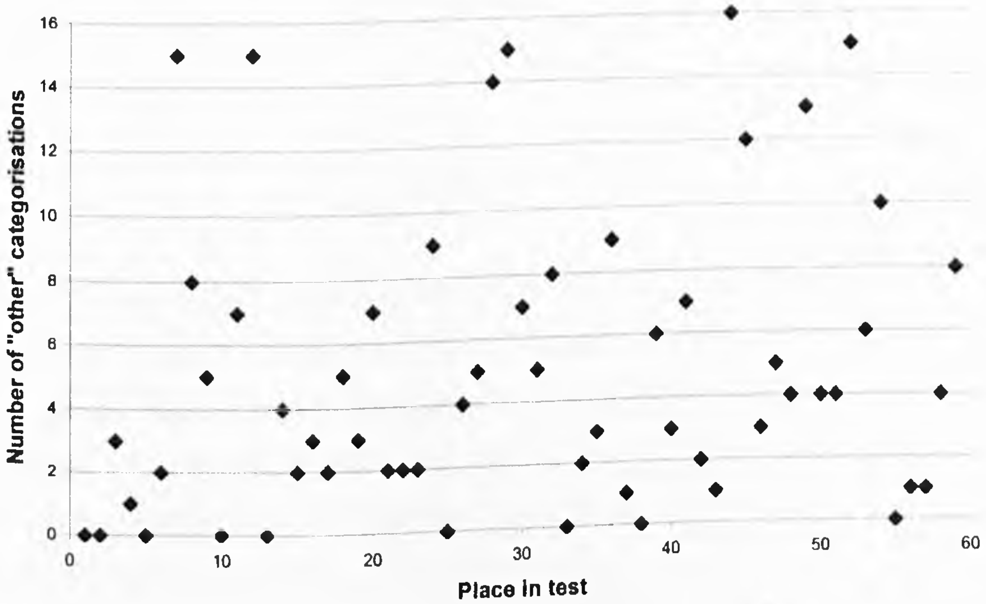


Figure 4.6 – Number of subjects categorising adjectives as "other" against place in test

Did lower placing in the categorisation experiment result in an increase in subjects using the "other" category? Figure 4.6 suggests not, although this does not show how many people suggested an alternative category. However, examination of subjects' responses shows that whether a subject gave an alternative category when using the "other" option was not related to the position in the test. Figure 4.6 shows that there was not a relationship between the frequency of word usage in the adjective gathering experiment and the ability of subjects to place that word in one of the categories given.

Figure 4.5 suggested that frequency of use and usefulness may not necessarily be related. Figure 4.7 illustrates this further, plotting the twelve adjectives rated worst for usefulness, along with the number of times they were used in the first experiment.

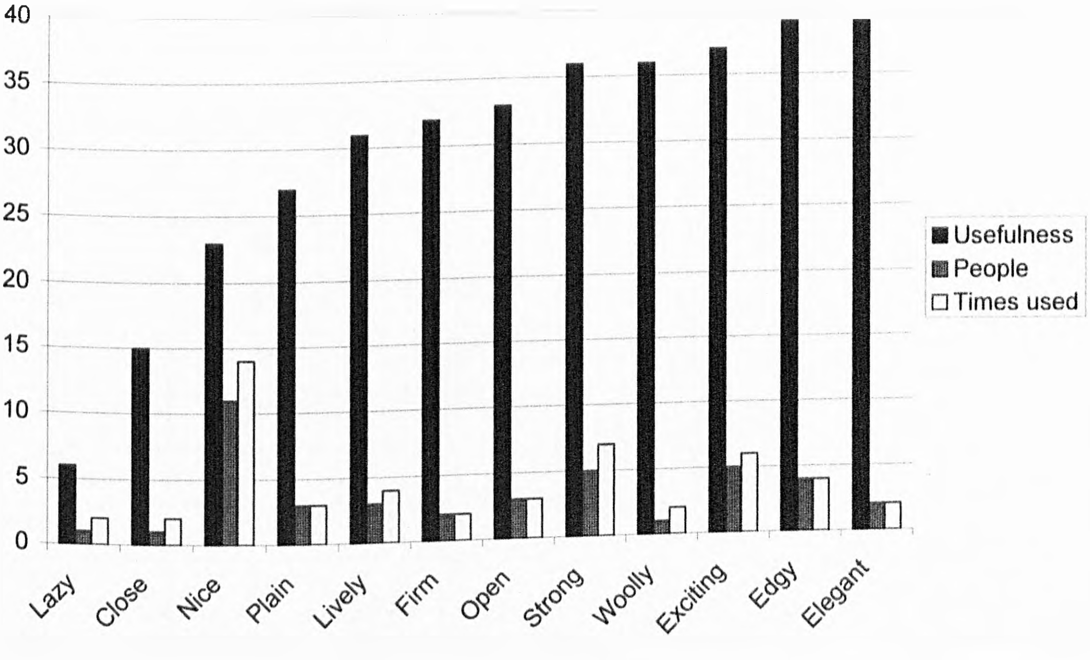


Figure 4.7 – Usefulness ratings compared with adjective gathering occurrence

Of these twelve words, only the first six were rated as useful by less than half of the 66 subjects taking part. The first two, "lazy" and "close", were two of the four words included in the adjective classification in error, as only one subject in the adjective gathering experiment used them. The other two such words were "woolly", which also appears in figure 4.7, and "shrill", which with "reedy" was joint fifth most useful word with a usefulness rating of 61. This concurs with the unpredictability of usefulness ratings of such less-used words suggested by figure 4.5.

Table 4.13 shows the top two adjectives from each category. What is clear from this and table 4.12 is that many people used "sound" more than any other category. In the cases of both "Vision" and "Taste & Smell" categories' top rated adjectives, there were still more who thought the words were sound-related in origin. Overall, the average number of times "sound" was given as a category was 30.47, with a standard deviation of 11.78. The lowest number of times "sound" was used as a category was ten, and the highest was 55, with six subjects using it fifty or more times.

Adjective	Sound	Vision	Emotion	Taste & Smell	Touch
Flutey	**66**	1	0	0	0
Nasal	**65**	0	0	1	1
Bright	38	**31**	0	0	0
Clear	38	**29**	1	0	1
Exciting	8	3	**55**	2	2
Nice	8	4	**51**	3	2
Sweet	26	2	20	**21**	2
Rich	42	4	11	**14**	1
Firm	11	1	9	1	**45**
Gentle	18	2	34	2	**43**

Table 4.13 – Highest scoring adjectives in each category

Subject comments provide some insight into the popularity of sound as a category. The following comment was typical of those subjects that made specific reference to their choice of category, and indicates the difficulty subjects had in considering the words both within and outside the context of the pipe organ.

> "I guess that I just equate most words as sound when evaluating the sound of an organ."

Clearly there is a bias towards sound as a category that renders the relative ratings in this section unreliable as precise categorisations, due to confusion between word origins and word usage. It is impossible to quantify that bias, although the results of this categorisation do show distinct trends for a number of words.

This use of subject-led adjective categorisation has shown that any such future experiments must clearly decide and communicate with the subjects the grounds upon which they want the words categorising. Such grounds could be the origin, usage, or applicability of such words. Any other parts of the test must not confuse the subjects on this, as it appears the usefulness ratings have done in this experiment, although they are potentially the more useful of the two in the wider context of this thesis. Unless future categorisation experiments are carefully designed in this manner, they may prove no more valid than the etymological approach this experiment sought to avoid. Authors must beware of putting their own personal preferences ahead of a rigorous approach to categorisation – as table 4.11 shows, the author of this report was often at odds in his opinion of a category with the majority of subjects.

4.7 Choice of words for further study

The criteria for selecting which words to study further are simple: the adjectives must be used, useful, simple and unambiguous in meaning. They should have been used by multiple listeners in the adjective gathering experiment, and also be considered useful by subjects from the categorisation experiment in the context of describing pipe organ sound. There should also be a limited number of words in order to facilitate detailed study of each word.

Some of the words fulfilling both those categories are more analytical or stylistic in nature, such as "neo-Baroque". These do not normally refer directly to the timbre but rely on a listener's analysis of the sound's likely historical background. Other words are ambiguous in that they can be used to describe several different aspects of pipe organ sound, and thus different listeners could put a different interpretation on that word. That could occur with any word, but where there is a clear indication that this is likely to happen, it makes sense to avoid the word in question.

An example of this potential dual meaning is "articulate". This could refer to the builder's voicing of an ensemble or to the touch and playing style of the performer. Clearly the latter is outside the scope of this study, but as previous experiments have shown that some subjects are distracted from their core instructions by other aspects of pipe organ performance and instruments, it is possible that some subjects would interpret "articulate" as referring to playing technique. Words used to describe the sound of an instrument might also be used by players to describe the feeling of playing an instrument.

Rioux and Västfäll (2001b) presented listeners with 85 adjectives gathered from experiments based on recordings of an individual organ pipe. Although a small number of samples were used, the length of the tests would have risked experimental fatigue even with the single examples presented to their subjects. Use of a similar number of adjectives with comparison tests would place too much stress on subjects. Rioux and Västfäll were only able to do general analysis of word categories as they lacked the time to specifically look at each word.

Although Rioux and Västfäll did reduce the number of adjectives (from 220 gathered), the nature of this reduction was not entirely clear, but involved removal of opined synonyms, uncategorised words, and some element of frequency based rejection. As they did not publish precise details of this process, it is impossible to comment precisely on the list of adjectives given, but it is likely that some of the adjectives presented to their subjects would have failed the criteria used by the author in this section. Certainly some of the words presented by them to subjects fall well beneath average usefulness in table 4.12, and it is preferable to exclude less useful words to avoid unnecessarily lengthy tests.

Creasey (1998, pp. 76-80) lists twenty-seven timbral descriptors for which some research has been previously conducted relating mostly single adjectives, although some pairs are included, to spectral features. It is interesting that the majority of these words have occurred in the unforced adjective gathering and categorisation experiments earlier in this chapter. Those twenty-seven words are listed in table 4.14 below, along with their various scores in this chapter's experiments. "L" indicates word occurrence information derived from the later categorisation experiment. Words in bold were included in the list presented to subjects in Rioux and Västfäll's later experiment (2001b).

Adjective	Number of subjects who used the word	Number of times word was used	Usefulness score
Acute	0	0	Not tested
Bright	26	41	64
Clang	0	0	Not tested
Clear	15	19	60
Cutting	1	1	Not tested
Dull	3	3	55
Fat	0	0	Not tested
Full	9	11	Not tested
Grating	1 (+L1)	1 (+L1)	Not tested
Hard	1 (+L1)	1 (+L1)	Not tested
Heavy	2	2	48
Hollow	2	2	54
Intense	0	0	Not tested
Jarring	0 (+L1)	0 (+L1)	Not tested
Lax	0	0	Not tested
Nasal	2	2	55
Open	3	3	33
Penetrating	0 (+L1)	0 (+L1)	Not tested
Presence	0 (+L1)	0 (+L1)	Not tested
Rich	7	8	55
Reedy	7	8	61
Resonant	3	3	53
Rough	1	1	Not tested
Small	2	2	Not tested
Soft	3	3	54
Thin	12	14	59
Warm	5	9	59

Table 4.14 – Adjectives previously related to spectral features (from Creasey)

Of the thirteen words in table 4.14 that were tested for usefulness, two (Heavy and Open) fell beneath a usefulness rating of 75% in the context of the pipe organ. Only five of the twenty-seven words did not occur anywhere in the experiments in this chapter. It would be very surprising if all those adjectives had occurred, as their various studies had been drawn from a wide range of fields. Indeed, as Creasey says (1998, p80):

> "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."

Although research may already have been carried out on some of the clear candidates for further study, it is not inappropriate to choose words that have been studied in existing work. None of the words have been studied in this timbre space, and they may have different understandings in this context. The previous work could be of help in providing a starting point for possible spectral correlates. It should be noted that Creasey did not go on to conduct any experiments directly relating timbral semantics and spectral features.

The adjective usefulness ratings were not available when the words for further study were selected, so selection was based on the frequency of occurrence in the adjective gathering experiment. Figure 4.8 shows the eleven most frequently occurring words (those occurring nine or more times) shown in figure 4.5 along with their usefulness ratings for comparison.

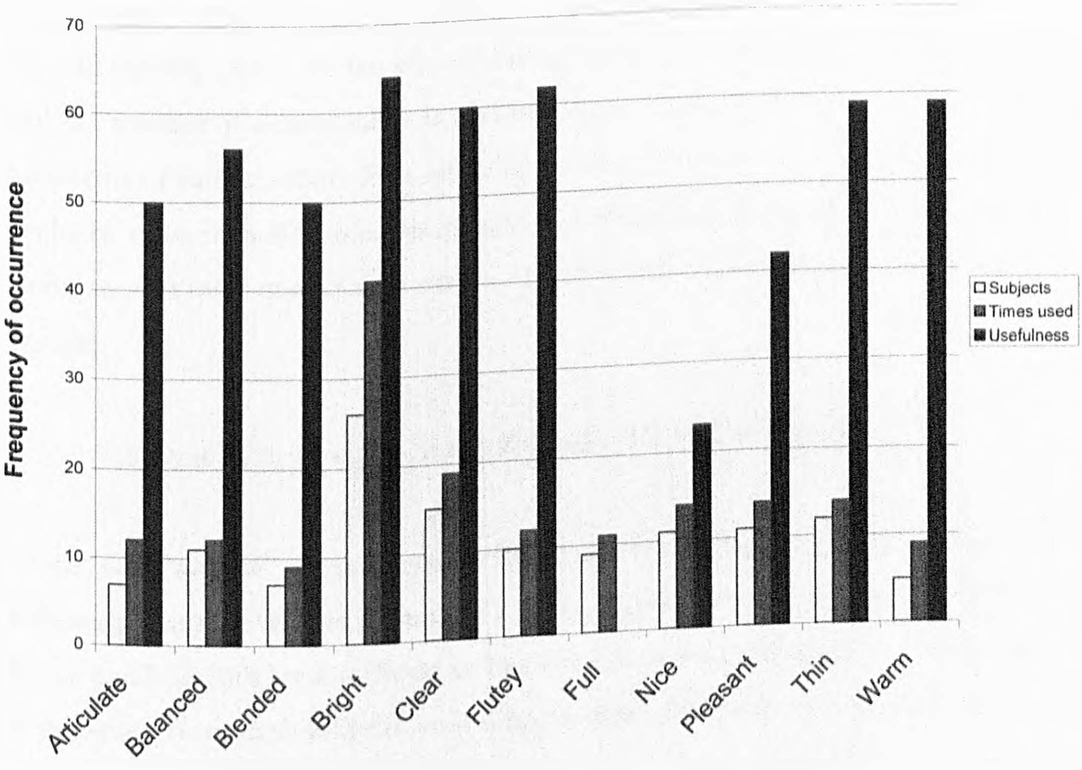


Figure 4.8 – The eleven most frequent words gathered with usefulness ratings

Of these eleven words, "nice" and "pleasant" clearly had more to do with personal preference than timbral correlates. In previous research, the author had already concluded that what makes an organ good or bad was an analytical process involving a combination of taste and many complex interacting timbral and other factors. The author has examined "blend" in some depth, including the introductory experiment in chapter three where a degree of common understanding was discovered.

"Articulate" was inappropriate for further study in the context of timbral adjectives relating to the pipe organ as it has the potential dual meaning described earlier in this section. This leaves seven words for further study. With the exception of "full", mistakenly left out of the experiment in which usefulness was rated, these words also score the highest usefulness ratings of all eleven words in figure 4.8. This backs up the author's exclusion of the other four words. The seven words chosen for further study all had usefulness ratings of 56 (84.8% of subjects) or over. Those adjectives were:

Bright, clear, thin, balanced, flutey, full, and warm.

All these words also occurred in the list of 85 adjectives given to subjects by Rioux and Västfäll (2001b). With the benefit of hindsight and the additional data on usefulness and the number of unique subjects who used each word, several other words might also be worthy of further study. Five adjectives were used by at least five subjects and rated useful by more than 80% of subjects, and also fulfilled the categories of not being ambiguous or more analytical in nature. These words were (with their usefulness ratings):

Chiffy (63), mellow (57), reedy (61), rich (55), and stringy (58).

"Rich" and "mellow" did not occur in Rioux and Västfäll's list of 85 words, possibly indicating that they refer to ensembles and not single pipes. However, comparison with Rioux and Västfäll's list is difficult as they did not collect geographical and language information from their subjects, so it is impossible to know to whom their results apply.

4.8 Conclusion

Fifty subjects contributed to the collection of 99 adjectives related to the timbre space of the pipe organ. Comparison with other studies and ratings of usefulness suggested a common core of timbral adjectives. However, there were some noticeable omissions when the commonest words were compared against scales commonly used by psychoacousticians, perhaps because the tester rather than the listener defines those scales.

The exclusion of inappropriate words and analysis of how frequently they occurred resulted in the choice of seven adjectives for further study. Subsequent measures of usefulness confirmed this choice.

5. Spectral correlation of timbral adjectives

In previous chapters, timbral adjectives were gathered and refined to produce a subset of commonly used timbral descriptors that may have common understanding through virtue of their frequent use and considered usefulness. This chapter examines existing studies of these words in different timbre spaces to suggest potential acoustic qualities they may refer to. In two experiments guided by these factors, common understanding and applicability of these theories to the timbre space of the pipe organ is tested. The significance of subject musicality is also tested.

5.1 Possible acoustic triggers for adjectives gathered

Having gathered suitable adjectives to study, it is proposed to use further musical examples within the timbre space of the pipe organ to examine common understanding. These musical examples must vary sufficiently to allow subjects to exhibit any such common understanding. It is sensible therefore to examine potential relationships between the words gathered and acoustic factors, particularly where work has already been done in other timbre spaces. This will permit educated selection of musical examples for use in comparison tests.

The words selected for further study in section 4.7 were:

bright, clear, thin, balanced, flutey, full, warm

All but "flutey" and "balanced" have relevant spectral characteristics in previous studies. Most of these are summarised in Creasey (1998, pp. 76-80).

Bright was described by Creasey as "the most common description to occur in the literature". Two alternative spectral effects have been associated with perceived brightness: spectral width and spectral centroid. Spectral width refers to the spread of frequencies from 0Hz upward, and the spectral centroid is the mean of a given spectrum.

The first theory is supported by research in Campbell (1994) and described in Dodge and Jerse (1985). The second is mentioned in Risset and Wessel (1999, p115), and supported by research in von Bismarck (1974 (1)). Von Bismarck (1974 (2)) states that the second means of achieving brightness may be related to "sharpness" and is perceptually less pleasant than the first method. Both theories could be difficult to isolate from each other, as conventional pipe organ ensembles always have a fundamental component and thus an ensemble with greater spectral width will have a higher spectral centroid. Some of Creasey's summary is for the word "brilliant", which has not been included here as a synonym due to its high individual usefulness score in chapter four (63/66, or 95.4%).

Creasey summarises research on "clear" thus (with references inserted):

> " Increasing clarity may arise from decreased number of harmonics, expanded harmonic density and decreased spectral slope (Ethington and Punch, 1994). Jeans (1961) believed that the 2nd harmonic's amplitude relates to clarity as it is an octave doubling effect. In Solomon (1959), clarity positively correlated with high frequency energy, neutrally with mid frequencies, and negatively with low frequency emphasis."

The theory of Jeans is interesting, as the addition of an octave harmonic (for example adding a 4' stop to an 8' stop) is a basic part of an organist's ensemble synthesis. The other theories can be compared with initial comparative test results to see if they are relevant in this timbre space.

Hall (1991) suggests that a perception of "thin" can be due to a lack of lower frequency spectral components. Creasey (1998, p80) gives two references from synthesiser users that appear to support this theory. This is a relatively simple spectral characteristic to aid selection of audio examples. Creasey also gives the following information on "full":

> "This is attributed to various effects, from having both odd and even harmonics (Pierce, 1983) to having more low frequency presence."

Of all the terms, "full" is least likely to match its general usage when it describes a pipe organ. "Full organ" is a term in common usage to describe the sound of all chorus stops on a particular instrument being played together. However, "full" lacks the obvious duplicity of meaning of, for example, "articulate", and this likelihood of specific meaning within the pipe organ community does not exclude it from further study. This lack of a clear theory means that "full" cannot help selection of initial examples.

Warmth is another perceptual quality with two suggested spectral associates. Risset and Matthews (1969), Pierce (1983), and Rossing (1990) all suggest that a small degree of inharmonicity of components nominally at harmonic multiples can produce a perception of warmth. However if this variation results in more than one component at a single harmonic multiple, the sound may be perceived as "rough".

The second potential spectral correlation is with a decrease in the number of harmonics and an increase in energy at lower frequencies, particularly the first three harmonics. Pratt and Doak (1976) presents this theory, and Ethington and Punch (1994) support it, also attributing perceived warmth to compressed harmonic density and decreased spectral slope.

Variation in harmonicity of spectral components is difficult to detect in the complex ensembles used in this research. However, the slight inharmonicity of partials identified as a causal factor in perception of "warmth" above cannot be applicable in the case of the pipe organ without causing the roughness also suggested. This is because each harmonic of each rank in a pipe organ must blend with the other ranks of the same pitch and at harmonic multiples thereof. Deliberate mistuning is rarely found in the pipe organ, and usually only in the case of entire ranks (e.g. Celestes) that produce an undulating effect that is not part of the principal chorus. This inharmonicity is not to be confused with research that suggests organ pipes may have natural resonances at inharmonic pitches (e.g. Kob, 2000). Such resonances only serve to modify the spectral characteristics of harmonics, and do not promote generation of inharmonic spectral components.

Much previous work on single source sounds cannot be directly applied to this project because of the multiple source nature of the ensembles being studied. However, the more general spectral theories are of use when selecting samples to aid further study.

Section 4.7 suggested five other words that may be worthy of further study in future experiments outside of this thesis:

chiffy (63), mellow (57), reedy (61), rich (55), stringy (58)

The author has been unable to find relevant previous research on the terms "chiffy", "mellow" and "stringy", although "chiffy" and "stringy" have particular meanings within the context of the pipe organ. Howard and Angus (2001, p222) suggest that a "reedy" quality is particularly dependent upon the seventh harmonic and to a lesser extent the fifth harmonic. Informal tests by the author confirm this perception, and suggest that this term might be particularly illuminating in the context of the pipe organ.

Creasey (1998) has the following to say about the term "rich":

> "This is considered to occur from having the first harmonics ascendent (up to about the 6th possibly) in von Helmholtz (1877). From the results of Seashore (1938), Solomon (1959) and Pratt and Doak (1976), richness/colour seems to relate to the number of significant harmonics. The 5th harmonic increases richness and an horn-like quality (Jeans, 1961)." [sic]

This last comment is as applicable to "reedy" as "rich", and the term "significant harmonics" needs further development, but both suggest potential for future study.

None of the theories examined in this section refer in any depth to spectral components of the transient portions of sounds. As spectra often change dramatically in this period, perhaps this is not surprising. Within the context of a large ensemble of pipes, all of which have their own transient, the concept of an ensemble transient is not one that lends itself to analysis. However, it is possible that, both individually and collectively, pipe transients have an effect on overall perceived timbre and timbral qualities.

5.2 Development of experiments

It is proposed to use the analysis and synthesis methodology described by Gabrielsson (1985) and explored by Risset and Wessel (1999) to explore these timbral descriptors further. This method selects and analyses acoustic examples, and isolates variables for further study. Synthesis then provides a number of systematically varying versions, which are then judged by listeners on a number of scales. Both listeners and their responses are categorised and studied in an attempt to obtain meaningful results.

This method is particularly suitable as it uses recordings in initial stages, avoiding the secondary nature of electronically generated sound and also satisfying the criteria developed in response to earlier criticisms of artificiality.

The suitability of comparison tests for this work has been previously described. Kostek (2003) rejected this approach due to the large number of audio samples he used and the additional time comparison tests would take. Rioux (2001) presented subjects with a reference tone in each example, eliminating the need to compare each of the other examples with all samples. However, this is only suitable where a reference point is known.

With seven adjectives, five of which had existing theories of acoustic correlation, it would have been difficult to adequately address all of them in one experiment along the lines of the preliminary work in chapter three. The opportunity arose to test a large number of people in the context of a seminar on this work. As these people were drawn from across and beyond the Department of Electronics at the University of York, their musicality would vary and provide an opportunity to examine how this affected their answers. However, as the seminar was on this research, hopefully sufficient numbers of people would attend who were musical or interested in the pipe organ, and thus provide a suitable subject group for this study.

Two words, "bright" and "clear" were chosen for this study. As these descriptors were most popular in the adjective gathering study, they had most chance of common understanding in the wider community. Both also had considerable previous research and relatively clear theories of spectral correlation, making them more suitable for a less focused subject group such as this. Fewer tests would be necessary to indicate potential compliance with these theories. Subsequent testing using synthesised examples would be no less rigorous.

The remaining five adjectives, with the reduced number of spectral correlation theories, could then be used in an Internet-based comparison test with specialised listeners. All words from both experiments that appeared to have common understanding will then be compared against spectral analyses to determine whether they have any identifiable correlation. This will permit the careful synthesis of examples to be presented to another specialised group.

5.3 Testing of general listeners

The theories for "bright" and "clear" have been presented in section 5.1. The samples use in this part of the research should begin to examine those theories in the timbre space of the pipe organ while both being suitable and taking advantage of the different circumstances of this particular test.

This test was given to an audience who attended a seminar on the subject of the author's research, and were asked if they would stay and take part in a brief experiment connected with the research. Approximately 90% of the audience chose to actively take part, and all chose to remain in the lecture theatre for the duration of the experiment.

The direct presentation of samples removed the restriction of file size that applied to Internet based tests. However, given that the audience was less specialised and there was a finite amount of time available, the overall test had to be shorter in length and succinct in what it asked of subjects.

The lecture hall sound system utilised for presentation of audio examples to the audience was stereophonic, of high quality and recently installed. As it was impractical to consider the precise frequency response available from each loudspeaker for each listening position, only comparison tests were appropriate. As a result of these considerations, it was decided to undertake two comparison tests, each using a different pair of samples. The samples themselves were of longer length, between 15 and 18 seconds, and replayed at the original 16-bit 44.1kHz stereo resolution. Careful choice of samples and the lack of cross-comparison meant that no pitch correction was necessary. The musical example used was the first two stanzas of "Old Hundredth".

As "clear" had several contrasting theories from different fields, the primary aim when selecting samples was to determine the factors behind perceived brightness. If an organist was asked to increase the brightness of an ensemble, he or she might add a mixture, a collection of pipes speaking at higher harmonics. The first two samples chosen were therefore of the same organ, with and without its mixture. The second two samples were of different organs, each with a mixture present but with their individual acoustic signatures. One had more isolated strong harmonics (including the second) and the other was more even in its harmonic development. The amplitude and frequency data for these ensembles can be found in appendix G.

The first test used two samples from the organ of St. Chad's. The first recording included principals at 8, 4 and 2 foot pitch. The second recording added a mixture including the fourth, fifth and sixth harmonics (a 15.17.19 mixture). This is comparatively unusual for a principal mixture, as the harmonics are usually found more spread out. Odd numbered harmonics are also unusual, as they alter the perceived timbre, particularly the fifth and seventh harmonics (see "reedy" in section 5.1 above). The aim here is to determine whether adding this more complex mixture still increases perceived brightness as might be expected from a simple study of existing theories, or whether the more complex nature of these added harmonics alters subject reactions. This may also indicate which of the two apparently conflicting theories (Ethington and Punch or Solomon, summarised in section 5.1 above) is likely to be applicable in this timbre space.

The second pair of samples used were the organs of the Jack Lyons concert hall and Doncaster Parish Church. Both were registered to their 8-foot mixture stop, with no 16-foot components. 16-foot or sub-octave stops are used in larger organs to create a more substantial sound, and the largest organs such as Doncaster Parish Church include stops such as a "Quint 5 1/3-foot" and mixture stops that have harmonic components from the 16-foot series. Previous inclusion of sub-octave components and their harmonics in research by the author (Disley, 2000) produced confusion in subsequent analysis, so only unison choruses have been examined in this instance.

The Jack Lyons Hauptwerk was registered with 8 and 4 foot principals, its 2 foot Flachflöte and its 15.19.22.26.29 Mixture V. The Doncaster Parish Church Great was registered with principals at 8, 4, 2 2/3 and 2 foot, including both at the unison pitch. The Cymbale mixture, composition 19.22.26.29, was added, as the Mixture on that manual broke back to include harmonics of the sub-octave on higher notes.

These ensembles were chosen as although they both have similar theoretical harmonic make-up, in practice their sound is very different, due to both the acoustics they are in and the tonal voicing achieved by their builders. Their comparison should shed light on both "bright" and "clear" theories. Specifications and overviews of all these organs can be found in appendix E.

Subjects were asked to compare each pair of samples on three scales:

- Brightness
- Clarity
- Preference.

The latter scale was included to give insight into later consideration of subject musicality and its influence on results. Scales were presented as continuous lines upon which subjects could make a mark:

Example 1 - - - - - - - - - - - - - - - - -✘ - - - - - - - - - - - - Example 2

This was used in preference to a quantised scale as it was considered to be simpler to explain and more intuitive to use. The results were later quantised to the eleven point scale used elsewhere in this research by a simple "nearest match" procedure.

Subjects were also asked to rate their own musicality to permit analysis of the results by musical ability. Given the imprecise nature of self-selecting musicality noted by the author in section 3.3 and the likely audience, a simple three-point scale was used:

- No musical ability
- Some musical ability (e.g. Grade V)
- Good musical ability (e.g. Grade VIII)

NOTE: *The suggested grade level refers to musical examination levels of the Associated Board of the Royal School of Music (ABRSM), which are in common usage in the United Kingdom and elsewhere. These range from elementary (Grade I) to accomplished (Grade VIII). Most professional music teachers encourage students to take such tests, and anecdotal evidence suggests that most students who do not go on to become professional musicians cease their formal musical education at either Grade V or Grade VIII. These points provide useful references that will have been known by most subjects to the otherwise vague terms "some" and "good".*

Subjects were also asked whether they had significant experience of either listening to or playing the pipe organ.

5.4 Results from general listeners

The overall results of the first test, comparing the two samples of St Chad's respectively without and with its mixture stop, are presented in table 5.1. The eleven-point scale is identical to that used in chapter three, and has a mean of six with extremes of one and eleven. Conversion of answers from this scale to percentages is also as described in chapter three.

Scale	Average	Standard Deviation
Brightness	6.82 (16.4%)	2.34 (46.8%)
Clarity	6.28 (5.6%)	2.27 (45.4%)
Preference	5.10 (-18%)	2.92 (58.4%)

Table 5.1 – Results for first general listener test

Those results are for all 75 subjects combined, with no consideration of subjects' musicality, which will be considered in section 5.4.1. In general, this appears to support the theories of brightness. It is interesting that the sound of the ensemble with this more harmonically complex mixture is perceived to be clearer, although only by a small margin. The preference rating shows the highest standard deviation, but at least suggests that these subjects did not have an overall bias towards the second sample – or indeed just the right hand side of the answer sheet.

The results for the second test, which compared (respectively) the Jack Lyons Hauptwerk ensemble and the Doncaster Parish Church 8' Great chorus, are presented in table 5.2.

Scale	Average	Standard Deviation
Brightness	7.11 (22.2%)	2.39 (47.8%)
Clarity	7.25 (25%)	2.44 (48.8%)
Preference	7.55 (31%)	2.56 (51.2%)

Table 5.2 – Results for second general listener test

Here all three results are similar in both average and standard deviation. It will be interesting to examine the correlation between subjects' answers to see if this was the case for individual subjects. The previous test suggests that subjects were capable of isolating each scale, so it should not be assumed that "brightness" and "clarity" were assumed to be positive characteristics inherent in a preferred example.

In all cases the standard deviations were comparable with the much smaller subjects groups of the initial experiments described in section 3.6. It will be interesting to examine how the standard deviation varies with subject musicality. In this limited example, the results appear to indicate a degree of broad common understanding, and this will be examined in light of subject musicality ratings in section 5.4.3.

5.4.1 The significance of musicality

Results divided by self-selected subject musicality are presented in tables 5.3 and 5.4. In all cases, subject numbers were well in excess of those given as minima by Levetin (1999). The smallest group, those listeners with regular experience of listening to or playing the pipe organ, was a combination of subsets from the other three main groups. This group comprised five of the "good" group, and four each of the other two groups. Their ratings individually did not have much effect on the total scores of their musicality groups, but when combined they demonstrate some distinct differences from the overall group trend.

Musicality rating (subject numbers)	Brighter		Clearer		Preferred	
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.
All subjects (75)	6.82	2.34	6.28	2.27	5.10	2.92
Good (19)	7.11	2.08	6.05	2.20	5.58	3.17
Some (22)	6.82	2.38	6.27	2.07	4.77	2.56
None (34)	6.53	2.49	6.50	2.46	4.94	3.03
Pipe organ experience (13)	7.15	2.97	6.54	2.79	4.15	3.18

Table 5.3 – Results for first general listener test divided by musicality

Musicality rating (subject numbers)	Brighter		Clearer		Preferred	
	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.
All subjects (75)	7.11	2.39	7.25	2.44	7.55	2.56
Good (19)	6.89	2.18	7.05	1.93	7.53	2.01
Some (22)	6.91	1.97	7.05	2.34	7.00	3.04
None (34)	7.53	2.73	7.65	2.77	8.12	2.46
Pipe organ experience (13)	5.38	2.69	7.08	2.81	6.69	3.12

Table 5.4 – Results for second general listener test divided by musicality

There were distinct differences between some of the groups in this experiment. When rating ensembles against adjective categories, the group who rated themselves "no musicality" had the highest standard deviation and thus the greater disagreement than the more musical groups. Those who rated their musicality "good" or "some" had, on average, similar results for both mean and standard deviation. This did not apply to the preference ratings for any group, in which the clearest trend to emerge was that of increased subject disagreement in all categories.

Also interesting is the subset of those who had experience listening to or playing the pipe organ. Usually this group had the highest standard deviation of all the subsets, suggesting greatest disagreement. The lowest standard deviation was for the "brightness" rating in the second test, where they were comparatively united in disagreement with the general trend. Only organists thought that the Jack Lyons instrument was brighter than Doncaster Parish Church, but with a standard deviation of 2.69 even this is not very conclusive. Dividing organists further by musicality results in group sizes too small to make meaningful comparisons from, but the data suggests that increased musicality might result in greater agreement.

Musicality does appear to be important, although the level of that musicality is less important. What is clear is that those with direct experience of the pipe organ in some circumstances have different interpretations of a word as it applies to the pipe organ. This supports the exclusive use of subjects with experience of the pipe organ in other tests, but also suggests that any spectral correlation of adjectives may not be relevant outside of pipe organ timbre space.

5.4.2 Covariance of results and variation with subject musicality

Whatever the test methodology chosen, it is wise to examine how test subjects are responding to and using the scales available to them. The preference result of the first test suggests that subjects weren't just selecting the same value on all three scales. To examine this more closely, the correlation between the various results needs to be examined. This correlation is between results for each individual subject, so although two subjects may give different ratings for the same sound, if they both prefer the example they rated as clearer, this will be identified by the covariance of the two scales.

Tables 5.5 and 5.6 present the Pearson correlation coefficients (introduced in more detail in section 3.7.2) for the first and second general listener tests. Table 5.7 looks for covariance between the same variable in both tests, to see if listeners' answers to one test were related to those of the other. Subjects were again split into the groups examined in section 5.4.1. Variables that meet statistical significance at the 5% level are marked in bold, and variables that meet statistical significance at the 0.5% level are marked in bold italics. The significance level varies with subject group size, and the values used here are taken from Howitt and Cramer (2000).

Subject group and musicality rating	Covariance between:		
	Brightness and clarity	Brightness and preference	Clarity and preference
All subjects (75)	-0.118	-0.016	**0.263**
Good (19)	-0.147	-0.178	**0.450**
Some (22)	-0.144	0.102	-0.149
None (34)	-0.079	-0.016	**0.373**
Pipe organ experience (13)	-0.474	0.094	0.375

Table 5.5 –Covariance for first general listener test divided by musicality

Subject group and musicality rating	Covariance between:		
	Brightness and clarity	Brightness and preference	Clarity and preference
All subjects (75)	***0.344***	**0.282**	***0.386***
Good (19)	0.252	***0.583***	**0.408**
Some (22)	**0.434**	0.222	***0.583***
None (34)	**0.321**	0.193	0.237
Pipe organ experience (13)	0.414	0.372	***0.943***

Table 5.6 –Covariance for second general listener test divided by musicality

Within the limited constraints of these tests, it appears that the preference of this subject group was related to the perceived clarity of the ensemble, particularly where organs in two different acoustic environments were presented. The preference of those with pipe organ experience for the clearer example in table 5.6 is very strong, although it is interesting to compare this with the standard deviations presented in table 5.4, which suggest disagreement as to which sample was clearer. The other strong preferences from table 5.6 are not matched in table 5.5, so it would not be appropriate to draw any other conclusions from this data.

Subject group and musicality rating	Covariance between:		
	Brightness	Clarity	Preference
All subjects (75)	-0.115	-0.175	**0.249**
Good (19)	-0.169	0.183	0.150
Some (22)	0.017	-0.248	0.251
None (34)	-0.136	-0.284	**0.318**
Pipe organ experience (13)	0.294	-0.165	0.014

Table 5.7 –Covariance between common variables of general listener tests divided by musicality

Table 5.7 suggests that subjects are, on the whole, able to exercise their judgement freely in the tests, and their answers to one test do not adversely affect subsequent answers. Given the relatively uncontrolled nature of the test and the wide variety of participant backgrounds, this is a useful suggestion that subjects' answers were not unduly compromised by test procedure. Neither of the two significant results in table 5.7 were highly significant, and some such limited covariance might be expected in any such real-world test. If subjects consistently showed such covariance over a test similar to that conducted in chapter three, it might indicate bias towards one answer. However, given that this test only included two sets of rating scales, this limited covariance does not present cause for concern.

5.4.3 Common understanding of timbral adjectives

As the standard deviations in tables 5.3 and 5.4 remain relatively high, it is difficult to draw specific conclusions, but some broad patterns appear evident. The results in those tables suggest that musicality plays a role in people's common understanding of timbral adjectives. As musicality decreases, disagreement between subjects increases, and yet their average answer moves further from the mean. This suggests that timbral adjectives have some degree of common understanding among subjects with all levels of musicality. Musical training has a clear effect, and exposure to a particular musical instrument (in this case the pipe organ) can alter people's usage of certain adjectives.

5.5 Testing of specialised listeners

The five remaining words selected (thin, balanced, flutey, full, and warm) were presented to a panel of specialised listeners as rating scales in a comparison test similar to that conducted in chapter three. Again, the Internet was used as a delivery mechanism, with a number of listeners taking part via the author's laboratory computer, using a pair of Sennheiser EH2270 headphones. No remote subjects reported any problems with test procedure, although to avoid one previous problem subjects used two different test pages differing only in the manner in which the sound file was linked.

The Internet subjects and most of the subjects who took the test locally were a subset of the group used in the adjective gathering experiment described in chapter four. Geographic data was collected from subjects, and age data was collected at a later date after repeated analysis of the initial experiment revealed possible age-related features. Sex data was not specifically collected, and informal evidence from subject names suggests that the female group would have been too small for meaningful comparison to be made. The total number of subjects was 22, and subject data is summarised in table 5.8. All subject group sizes, with the exception of Canadians, were sufficiently large to make later analysis appropriate following the size guidelines of Levetin (1999). The cross-referencing between the subject groups shows that the age groupings have a reasonably good geographic spread and vice versa, so none of the age/geographic groups is essentially made up of the same subjects.

Country of residence	Age group		Total
	Under 50	Over 50	
United Kingdom	6	3	9
United States	7	4	11
Canada	0	2	2
Total	13	9	22

Table 5.8 – Geographic and age data for specialised listeners

Most of the five words chosen for study in this experiment had less or no previous research into their spectral correlates, making better understanding of their acoustic triggers more important. However, even if a spectral correlate cannot be accurately identified, this does not preclude the possibility of demonstrable common understanding. Thus although potential spectral correlates can help in selection of examples, it is not necessary at this stage to rigorously adhere only to theories from previous research.

Of those theories described in section 5.1, the most clearly applicable to pipe organs is that for "thin". Hall (1991) has suggested that a perception of "thin" can be due to a lack of lower frequency spectral components. Therefore, some of the samples chosen should have varying degrees of lower harmonics. Although all of the principal ensembles being considered within the scope of this experiment are based upon an 8' component, the relative strength of this compared to the other stops used to build up the chorus can vary.

The second theory of warmth presented in section 5.1 (a decrease in the number of harmonics and an increase in energy at lower frequencies, particularly the first three harmonics) can also be applied relatively easily to pipe organ ensembles. Although an organist's natural attempt to achieve this might be to alter the stops drawn, perhaps removing a mixture such as that added in section 5.3 to create brightness, as with "thin" different organ ensembles will have differing degrees of lower and upper harmonic strength. These theories suggest that perceived warmth in particular might negatively correlate with perceived brightness or thinness, and hence brightness and thinness might positively correlate. Examination of the correlation between variables should prove whether this is the case. The final adjective for which there is previous research, "full", has no single clear theory, so if a consistent understanding emerges, it will be interesting to compare the two theories summarised in section 5.1 to see if either fits that word in this context.

Previous listener comments on the absence or artificiality of reverberation had emphasised the importance of presenting real-world samples with an adequate and genuine reverberant tail. Incorporating a reverberant portion is difficult to achieve with only a partial sample, but informal questioning of test subjects suggested that only about half would be prepared to download the file sizes necessary for longer tests. Remaining within the other constraints of sample size discussed in section 3.1.3 (in files of less than 1MB, at uncompressed 16 bit, 22.05kHz resolution) meant that only two lines of a hymn sample were possible. The only way to incorporate the reverberant tail was to present listeners with the last two lines of a hymn rather than the first two. A rapid fade-in would ensure minimal interference from the preceding two lines, but some reverberation from that would inevitably be present in the early part of the sample.

Similarly, about half of the subjects suggested that they would not want to take part in many more tests within the overall experiment. This concurs with the conclusions from section 3.7.3. From an experimental point of view, an increase in the number of tests would be beneficial. An odd number would permit presentation of each sample as the first and second of a pair in equal amounts. An increase in the number of samples in general would be likely to offer more insight into any theories developed. However, as an increase of tests by one sample would nearly double their number (ten instead of six) and potentially result in a significant loss of listeners, it was decided to remain with the four sample model previously used. The results in section 3.7.1 show that there was little bias despite the inequal distribution of each sample, but the answers to this part of the test must also be examined to ensure that any skew due to this has not occurred.

Listener comments in previous tests also suggested that the hymn tune previously used might be unsuitable for future tests due to overuse. In the sample library created by the author was a second hymn tune composed for the sole purpose of unfamiliarity. The third and fourth lines of that tune had the advantage of covering a melodic range of an octave in the treble, a tone more than the previous sample used. There was also an extract of a Bach fugue, but that was too lengthy for use and did not conveniently edit into a sufficiently small self-contained portion. The unfamiliar hymn tune was therefore used.

The question then arose of which registrations to present to the listeners. Given the specialised nature of listeners in this test, presenting recordings of obviously different registrations might have unwittingly suggested to subjects that a certain result was desired. Listeners may have used secondary reasoning, such as "One of this pair of samples does not have a mixture, therefore the less bright one is the one without the mixture", rather than their instinctive reaction.

As the principal ensemble of most of these instruments included a mixture, it was decided to select samples using the full principal chorus. Where that included sub-octave components, the full unison chorus omitting any sub-octave components was used instead. The mixtures add a degree of complexity for the proposed synthesis of later examples, but the synthesis engine was theoretically able to implement mixtures. Any consideration of a partial principal ensemble that could have included a mixture might not be directly comparable with another such ensemble, and in the case of the Jack Lyons sample, the mixture was a particularly vital component of the principal chorus.

There were five candidates in the sample library that included a mixture: Doncaster Parish Church, Heslington, Jack Lyons, Port Sunlight and St Chad's. Although Heslington's mixture had been added after the organ was originally built, this was done at the time the organ was installed in its current location and thus did not represent an alteration of the original builder's intention. The original recordings had been made by manually performing the piece of music as no reproduction devices were available on the organs recorded. The Doncaster Parish Church sample was slightly slower than the others, and the Port Sunlight sample was slightly faster. Use of both samples would result in too great a variation in sample speed, so as file size was also important, the Doncaster Parish Church sample was omitted. Port Sunlight was also the sample with greatest spectral potential for perceived "warmth". Acoustic analyses of all samples used will be considered in section 5.7, but the four remaining samples presented a good range of different principal ensemble examples that comprised roughly the same stops, including a mixture.

The precise samples used were:

- Heslington (Open Diapason 8', Principal 4', Fifteenth 2', Mixture III)
- Jack Lyons (Principal 8', Octave 4', Flachflöte 2', Mixture V)
- St. Chad's (Open Diapason 8', Principal 4', Fifteenth 2', Mixture III)
- Port Sunlight (Large Open Diapason 8', Principal 4', Fifteenth 2', Mixture III)

Ideally, tests would be presented in a random order to test subjects. This is difficult to implement over the Internet, and with such a small number of samples would have a high risk of repeating samples in consecutive comparisons. This can be avoided with a fixed test order. While everyone then has the samples presented in the same order, the effect of this order is reasonably predictable, and has been shown in section 3.7.1 not to unduly bias results. Samples were presented in the following order, again seeking to present each sample both first and second at least once, without playing the same sample twice in succession.

First sample – Second sample
Heslington – Jack Lyons
St Chad's – Port Sunlight
Heslington – St Chad's
Jack Lyons – Port Sunlight
St Chad's – Jack Lyons
Port Sunlight - Heslington

5.6 Results from specialised listeners

Listeners' results were analysed in the same way as those in chapter three, resulting in totals for each category and organ expressed as percentages of the theoretical maximum. The results themselves can be found in appendix H. Overall results for all 22 subjects are presented in table 5.9.

Organ	Thin	Flutey	Full	Warm	Balanced
Heslington	3.64%	-1.82%	1.52%	-4.85%	11.21%
Jack Lyons	26.67%	-32.73%	13.33%	-49.09%	-11.82%
St Chad's	-10.61%	-0.61%	-4.55%	20.91%	-2.12%
Port Sunlight	-19.70%	35.15%	-10.30%	33.03%	2.73%
Average standard deviation	2.12	1.99	2.59	1.80	2.30

Table 5.9 – Test results from specialised listeners

These results will be analysed in detail in following sections. One interesting thing to notice is that those adjectives with the highest standard deviation (full and balanced) are also those with the overall averages that deviate least from the centre. These together suggest that those words are the least commonly understood and the least useful in describing the differences between these particular samples.

Section 5.6.1 compares results between older and younger subjects, and those in different countries. Section 5.6.2 uses covariance testing to examine how subjects used the rating scales and whether any pairs of words had positive or negative covariance. Section 5.6.3 looks at what degree of common understanding is evident among the specialised listeners, and section 5.7 goes on to consider a number of spectral features with which these adjectives might correlate.

5.6.1 Significance of age and geography on subject results

Sections 3.6 and 3.7 suggested that there might be some significant age-related factors in people's use and definition of certain words. This specialised subject group also had sufficient listeners from both the United States of America and the United Kingdom to permit comparison of their answers. Results for the thirteen listeners under fifty years old at the time of taking the test are presented in table 5.10, and results for the nine listeners over fifty are presented in table 5.11. Results for the nine listeners from the UK are in table 5.12, and results for the eleven US listeners are in table 5.13.

Organ	Thin	Flutey	Full	Warm	Balanced
Heslington	1.03%	-1.54%	6.15%	-1.03%	21.03%
Jack Lyons	26.67%	-32.31%	17.44%	-54.36%	-30.26%
St Chad's	-9.74%	5.13%	-15.38%	17.44%	0.51%
Port Sunlight	-17.95%	28.72%	-8.21%	37.95%	8.72%

Table 5.10 – Test results from younger listeners

Organ	Thin	Flutey	Full	Warm	Balanced
Heslington	7.41%	-2.22%	-5.19%	-10.37%	-2.96%
Jack Lyons	26.67%	-33.33%	7.41%	-41.48%	14.81%
St Chad's	-11.85%	-8.89%	11.11%	25.93%	-5.93%
Port Sunlight	-22.22%	44.44%	-13.33%	25.93%	-5.93%

Table 5.11 – Test results from older listeners

Organ	Thin	Flutey	Full	Warm	Balanced
Heslington	13.33%	4.44%	1.48%	-5.93%	9.63%
Jack Lyons	13.33%	-22.22%	21.48%	-43.70%	-20.74%
St Chad's	-14.81%	-4.44%	-9.63%	28.15%	14.81%
Port Sunlight	-11.85%	22.22%	-13.33%	21.48%	-3.70%

Table 5.12 – Test results from UK listeners

Organ	Thin	Flutey	Full	Warm	Balanced
Heslington	-1.82%	-10.30%	0.00%	-3.03%	14.55%
Jack Lyons	35.15%	-43.64%	6.67%	-55.76%	-11.52%
St Chad's	-9.70%	7.27%	-4.24%	13.33%	-16.97%
Port Sunlight	-23.64%	46.67%	-2.42%	45.45%	13.94%

Table 5.13 – Test results from US listeners

When the results for young and old listeners are compared (tables 5.10 and 5.11), there are some marked agreements between the two groups of listeners. For the word "thin", the results are very closely matched, being identical in the case of the Jack Lyons. Results for the word "flutey" are also fairly closely matched, with both groups rating the Jack Lyons as "not flutey" to a very similar extent. Port Sunlight is considered strongly "flutey", although to different extents. Heslington is very near to the centre and St Chad's is not far away, with slight disagreement.

Both groups strongly agree on a definition of "warm", differing only by the degree of warmth. Port Sunlight was thought more warm than St Chad's by the younger subjects, but equally warm by the older subjects. However in the one comparison of Port Sunlight and St Chad's, both groups thought that Port Sunlight was more warm, the younger group by 12.3% and the older group by 7.8%. The remainder of their respective scores came from comparison with the other organs in the test.

"Balanced" appears to have poor common understanding, as does "full". Younger subjects think that the Jack Lyons is poorly balanced, but older subjects rates it as the best balanced of the four organs. However this is only by a small extent, and it must be remembered that the standard deviations for "balanced" and "full" were greatest of all the words being considered here.

On average, younger listeners tended to use the extremes of the scales available to them slightly more than older listeners, but not significantly so. The average absolute percentage value used by younger subjects was 17.08%, compared to 16.37% for older listeners. There was also a slight difference in the average standard deviation, suggesting that older subjects disagreed slightly more. The average standard deviation for younger listeners was 2.01 (40.2%), compared to 2.21 (44.2%) for the older age group. This compares with an average standard deviation for the entire specialised listener subject group of 2.16.

It is interesting to consider where this difference of age has come from. A strong possibility is subjects' differing hearing abilities, as the ability to hear high frequencies tend to deteriorate with age. Another possibility is the type of organ subjects are most familiar with, particularly those the subject was exposed to in formative years.

However, such speculation is not of great use without firm foundation, and is outside the scope of this thesis. What is clear is that dividing results by listener age results in significant differences for some words but not others.

Dividing listeners by geographic location also results in some significant differences as well as many similarities. United Kingdom listeners results are presented in table 5.12, and United States listeners are in table 5.13. Canadian listeners have not been included in either group.

For the adjective "thin", there was broad agreement as to the pairs that were and were not "thin", but there was disagreement as to the finer distinctions. United Kingdom listeners equated Heslington and the Jack Lyons, although in the direct comparison Heslington was thought "thinner" by 6.71%. United States listeners clearly thought the Jack Lyons was the "thinner" of the pair, by 11.81% in the direct comparison. Port Sunlight and St Chad's were both thought not "thin", but the percentages differ.

For "flutey" there was agreement on the extremes, but less agreement on the amount of the extremes and the order of the middle two. "Full" resulted in some very different answers, apart from agreement that the Jack Lyons was the most "full" of all the organs. United Kingdom listeners were far more definite about this than those from the United States, who barely moved from the middle. This suggests no great opinions either way, but this was more a result of differing opinions cancelling each other out. United States listeners had an average standard deviation of 2.72 (54.4%) for the "full" ratings, compared with a still-high 2.62 (52.4%) for the United Kingdom group

The word "warm" again has broad agreement. Both groups agree on the least warm by a big margin, only disagreeing on the order of the first and second. The different ratings for St Chad's and Port Sunlight are mirrored in their individual comparison tests. Finally, "balanced" again brings distinct disagreement. Although Heslington was thought "balanced" by both groups and the Jack Lyons was not, there was major disagreement on the other two organs.

The United States subject group made fuller use of the rating scales than the United Kingdom group, with an absolute average US answer of 18.3% compared with 15.04%

for the UK listeners. There was marginally less difference between the average standard deviations, with the UK averaging 2.18 (43.6%) and the US 2.11 (42.2%). As for where these differences came from, most of the significant differences appeared on words for which there was a large amount of disagreement among each subject group. Other differences could be due to different national organ styles, or even different national temperaments. Again, such speculation though interesting is outside the topic of this thesis.

5.6.2 Covariance and skew of results

Skew and covariance were introduced in sections 3.7.1 and 3.7.2. Skew is fairly easy to notice as it will result in an average result to one side of the central value of 6. In the case of the specialised listener test, the overall average answer was 5.96. When each subgroup's answers for each word were averaged, the largest deviation from the central value was 0.52 or 5.2%, well inside the 33% that would have suggested test order was having a significant effect on the results.

Covariance between the five words in this test would imply a link, either positive or negative, between their meanings. The covariance between each person's ratings was calculated, and the average Pearson correlation coefficient for each pair of words is presented in table 5.14. Some correlations for individual pairs of words were much higher than these figures, but that could simply be due to agreement on those two variables between listeners to that particular pair. Only by averaging the results can a true picture be created of the overall relationship between these adjectives.

	Thin	Flutey	Full	Warm	Balanced
Thin	(1)	0.022	-0.359	-0.265	-0.092
Flutey	0.022	(1)	-0.183	0.170	0.000
Full	-0.359	-0.183	(1)	0.342	0.270
Warm	-0.265	0.170	0.342	(1)	0.315
Balanced	-0.092	0.000	0.270	0.315	(1)

Table 5.14 – Average Pearson correlation coefficients

None of those correlation coefficients are significant at the 5% level for a group of 22 subjects. However, the negative relationship between "thin" and "full" is only 0.001 below significance. Other results below significance by a comparatively small margin include positive correlations between "balanced" and "warm", and "full" and "warm". These indicators of a relationship are particularly interesting given the large standard deviations on "full" and "balanced", suggesting that any such relationship might be as much due to a lack of clear understanding of "full" and "balanced" as any other factor. The possible negative correlation between "warm" and "thin" suggested in the previous section by previous theories is not at or near a level of statistical significance.

5.6.3 Common understanding of timbral adjectives

Of these five words, three ("thin", "flutey" and "warm") have clear common understanding demonstrated in table 5.9 by distinct high or low ratings for certain organs coupled with a relatively low standard deviation. The distinctly different answers of the various subject groupings for "balanced" suggest that it doesn't have any common understanding in this context. Similarly, despite more apparent agreement in table 5.12 from the United Kingdom sub-group, "full" lacks any real common understanding in this context as all groups had high standard deviations and the United States group was almost random in its ratings for this word. The differences due to geography appear to vary with some words, but this is mainly the precise order and score of the organs rated rather than overall differences. On the whole, for the three words identified, both age and geographic groups had similar understanding. We can quantify that similarity by looking at the correlation between both groupings for each word. Any significant correlation (at the 5% level for four samples) is marked in bold.

Subject subgroups	Thin	Flutey	Full	Warm	Balanced
Young and old	**0.981**	**0.942**	0.072	**0.959**	-0.864
US and UK	0.746	**0.924**	**0.966**	**0.919**	0.096

Table 5.15 –Pearson correlation coefficients between subject subgroups

The comparison of both sets of data allows clearer selection of the useful adjectives. "Full" scores highly in the US/UK comparison, but it must be remembered that the level of US response was much lower than the UK. Interestingly "balanced" is nearing significance with a negative correlation between young and old, signifying a tendency towards opposite understandings of that term. "Flutey" and "warm" both correlate significantly in both groups, and although "thin" is not so strong, a clear tendency towards correlation is visible in the insignificant example.

Rejection of adjectives without common understanding is necessary at this point, for without common understanding, no spectral correlation can be found and the words can serve no purpose in the remainder of this experiment. Rejection of "full" and "balanced" is valid on three grounds. Firstly, they have the least subject agreement, signified by their higher standard deviations than the other words. Secondly, they are the words on which there is most disagreement between subjects when divided by age and geography. All other words retain quantifiable common understanding between those groups – apparent differences are mainly of order and nuance, with clear patterns visible for certain words. Finally, both rejected words demonstrate a tendency to correlate with "warm", although not to levels of statistical significance. This tendency could be due to subject uncertainty leading to a subconscious partial correlation, or it could be due to an actual correlation in some subjects. In either case, apart from other considerations, the inclusion of known correlated scales in future testing or attempts to determine spectral correlates could lead to confusion.

It is important to note at this stage that the rejected words might still have meaning and common understanding between subjects outside of the constraints of this particular experiment. Their exclusion here does not preclude their use as descriptions of pipe organ timbre, but potential users should be aware that they are likely not to have common understanding and thus be a cause of confusion. It is entirely possible that some subjects were unable to divorce "full" from its meaning of "all stops playing", and that other subjects were describing the "balance" of different ensemble qualities.

5.7 Consideration and analysis of results

This section introduces a number of steady-state spectral analysis techniques that are then compared with subject ratings to look for correlation. The results of this will help to guide future synthesis of examples to verify both the common understanding exhibited by the results presented thus far and any theories of spectral correlation developed in this section.

5.7.1 Analysis techniques utilised

In the following sections, several analysis techniques are used that require introduction. Each real ensemble used in this research had single note samples taken from every C and F sharp on the keyboard for every combination of registration used. These were recorded from the same microphone position as the musical extracts, as to do otherwise would be to ignore the effect of the room upon the sound and listener. These samples were then analysed for harmonic amplitude information. This was done using a custom written Fourier based Matlab script, which can be found in appendix I.1. The script uses a 4096 point Hamming window to extract spectral data from a steady-state signal. The script then applies a simple peak detection algorithm to look for the maximum amplitude where each harmonic should be. This is limited to thirty-two harmonics to allow easy comparison of data across most of the keyboard range.

The amplitude of each harmonic is presented in decibels above the noise floor. An alternative method would be to set the amplitude of the most prominent harmonic as a zero point, as Guettler et al. (2003). However, because of the problem of accurately and consistently identifying the zero point, this is more suited for a highly controlled recording environment or, as in Guettler et al., synthesised examples. Beurmann and Schneider (2003) used the zero noise floor method when presenting waterfall diagrams of harmonic amplitude data, and it seems more appropriate for real-world use than the arbitrary zero method.

Visual presentation of this frequency amplitude data in this chapter uses the Acoustic Signature method developed by the author in Disley (1999). This presents the harmonic amplitude data with a third axis of keyboard range, resulting in a three-dimensional representation of the steady state harmonic content across the keyboard.

A number of different measures can be derived from this data. The spectral centroid is the mean point of energy distribution in the spectrum, and is normally measured in Hertz. For ease of comparison, it can also be normalised by dividing by the fundamental frequency. This means that other measures, such as consistency across the keyboard range, can easily be derived. The formula used for calculating the frequency normalised spectral centroid is:

$$f_{cN} = \frac{\sum_{n=1}^{32} f_n a_n}{f_1 \sum_{n=1}^{32} a_n}$$

where f_{cN} is the frequency normalised spectral centroid, f_1 is the fundamental frequency, and f_n and a_n are the frequency and amplitude for harmonic n. As the centroid is calculated for each of the sample points, two other useful measures can be derived: the average fall-off from the low to high extremes of the keyboard range, and the consistency between one note's centroid and the next. The former identifies how bright the lower notes are compared to the higher notes, and the latter is a measure of overall spectral consistency. The formulae used to calculate those measures are:

$$f_{cFALLOFF} = \frac{\sum_{1}^{\frac{S}{2}} f_c}{S/2} - \frac{\sum_{\frac{S}{2}+1}^{S} f_{cS}}{S/2} \qquad f_{cCONSISTENCY} = \frac{\sum_{n=2}^{n=S} f_{cn} - f_{c(n-1)}}{S-1}$$

where S is the number of samples. Some further measures are possible from the amplitude data. The average strength of all harmonics gives an indication of harmonic density in combination with centroid data, and is simply the mean strength of all harmonics for a given sample point. Using the same equation as for the spectral centroid, the fall-off of average harmonic strength over the keyboard range can be calculated and is a useful measure of a pipe organ ensemble. Organ builders sometimes make adjustments in pipe scaling to achieve certain desired emphases of strength in the treble. The average consistency of strength provides a more precise measure of how even the ensemble is from one note to the next, and is again calculated in the same manner as for the consistency of spectral centroid.

Spectral smoothness can be determined for a particular note by looking at the average difference between the strength of consecutive harmonics. As this measure doesn't always tail off towards the upper end of the keyboard range, the standard deviation of all smoothness values is a more useful measure of how consistent this is. Smoothness is calculated with the following equation (for 32 harmonics in this instance):

$$Smoothness = \frac{\sum_{n=1}^{n=31} \left| a_n - a_{n+1} \right|}{31}$$

This can then be averaged over the keyboard range to give a single value for each ensemble. A final measure derived from spectral analyses is that of spectral slope. This measures the degree to which the lower harmonics are greater in amplitude than the higher harmonics for any given note. This is achieved by averaging the harmonic strengths of each half of the spectrum and then subtracting the average of the higher harmonics from the average of the lower harmonics. Again, the standard deviation of all spectral slopes over the keyboard range measures the consistency of spectral slope. However, as most organ samples have a comparatively strong fundamental component, this measure is fairly crude. The inter-quartile spectral slope is similar in methodology, but excludes the lowest and highest quartiles, concentrating on the central portion where most of the change in amplitude occurs. Once again, its standard deviation measures consistency of this measure across the keyboard range.

The analyses described above are all calculated for the first thirty-two harmonics. Although some of the lower notes had visible harmonics above this, to include a variable upper limit would have meant the results were not comparable from one organ to another. In all cases this was limited to the lowest note samples and harmonic energy above the 32nd harmonic was less than 1% of the total spectral energy. The higher note samples had a different problem, as some of the 32 harmonics were above the Nyquist frequency. As these harmonics could not have been recorded, the highest three samples were limited to 24, 16 and 8 harmonics for analysis respectively. This was approximately 17kHz in each case, and the ability to divide each analysis into four quartiles was retained. Thus artificial padding of some formulae with zero harmonic amplitudes for those harmonics above the Nyquist frequency was avoided, resulting in more consistent performance in the various analysis methods used.

These methods for calculating spectral features do not include inharmonic data. This is because the ensemble of a pipe organ is entirely harmonic in nature. While recordings of individual pipes may exhibit inharmonic noise-like components such as wind noise and "chiff", particularly during the starting transients, such components have negligible spectral energy at listening positions. This does not mean that they are perceptually insignificant, but that in the context of steady-state analysis they are not readily isolated from background noise. All readily isolated spectral energy falls within the expected harmonic series. All the analyses described in this section derive entirely from steady state data. It is possible that study of transients using methods such as waterfall plots might give insights into subjects' perceptions, but it is difficult to apply traditional psychoacoustic measures such as DPAT (Duration of Perceived Attack Time) to such complex signal sources.

5.7.2 Spectral analyses of pipe organ ensembles

Full numeric data for all of the analyses in this section can be found in appendix G. The acoustic signatures of all six ensembles are presented in figures 5.1 to 5.6, after which the results of the analyses described in section 5.7.1 will be presented in a series of comparative graphs. Although the six ensembles analysed here were never all used in the same test, they are directly comparable, and the consideration of all six together will help in the identification of redundant potential spectral correlates.

The acoustic signatures were created in Microsoft Excel, using the 3-D Surface graph option. 3-D view parameters for all graphs were set to an elevation of 20, perspective of 30 and rotation of 15, and the colour scheme is the default setting. Using such an unspecific program can have disadvantages, which in this case manifests itself as a different depth to the Port Sunlight graph, despite all settings being identical. Despite those difficulties, this remains a simple and easily comprehensible presentation of amplitude data.

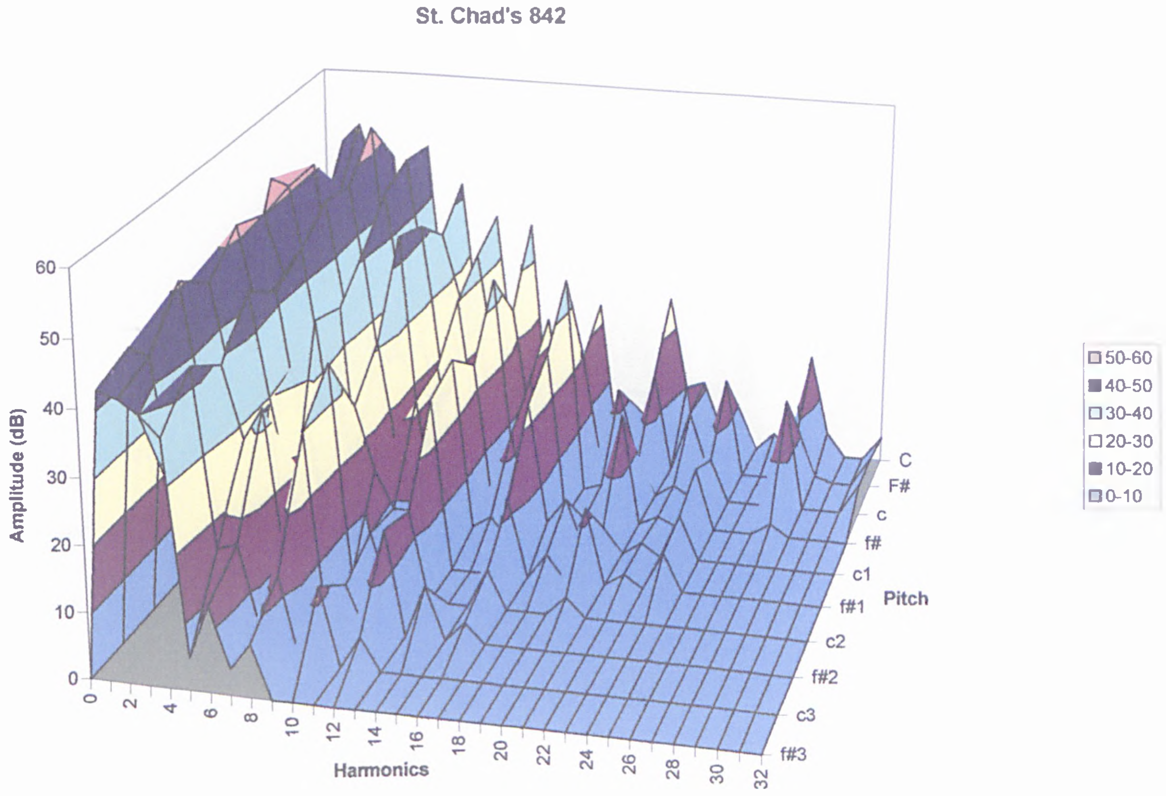


Figure 5.1 – Acoustic signature for the St Chad's ensemble without mixture

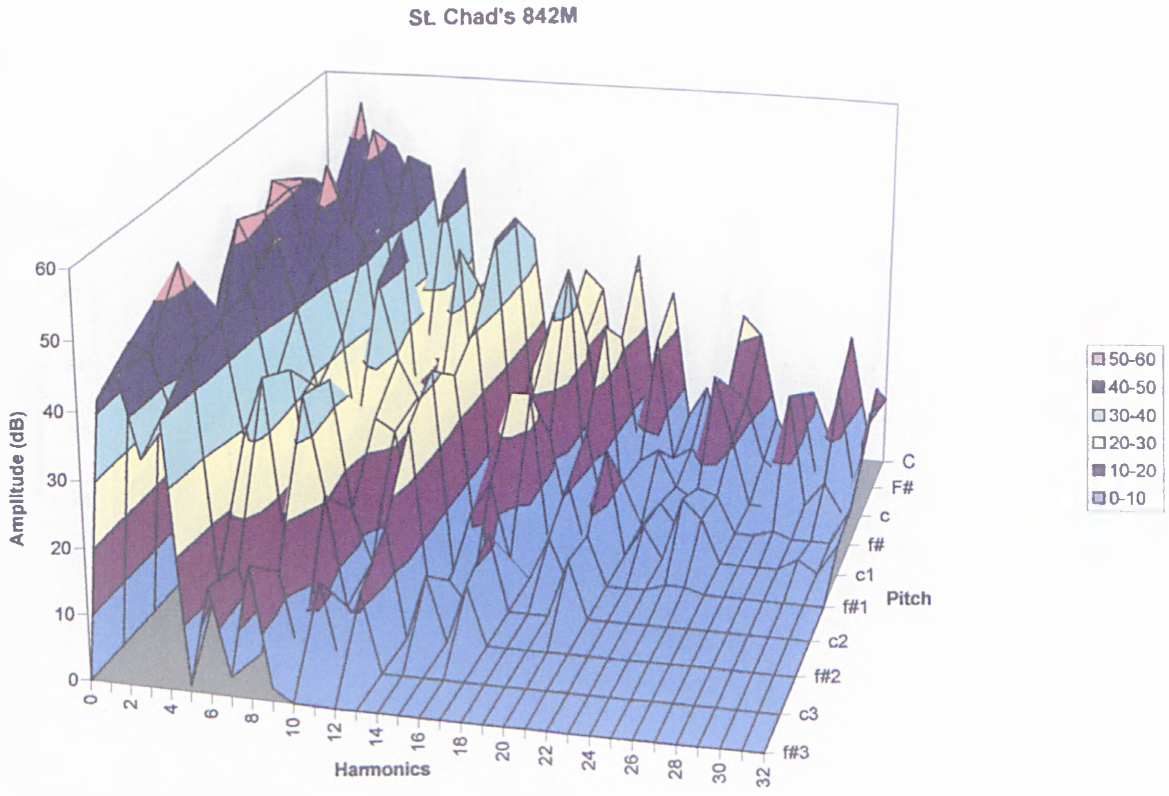


Figure 5.2 – Acoustic signature for the St Chad's ensemble with mixture

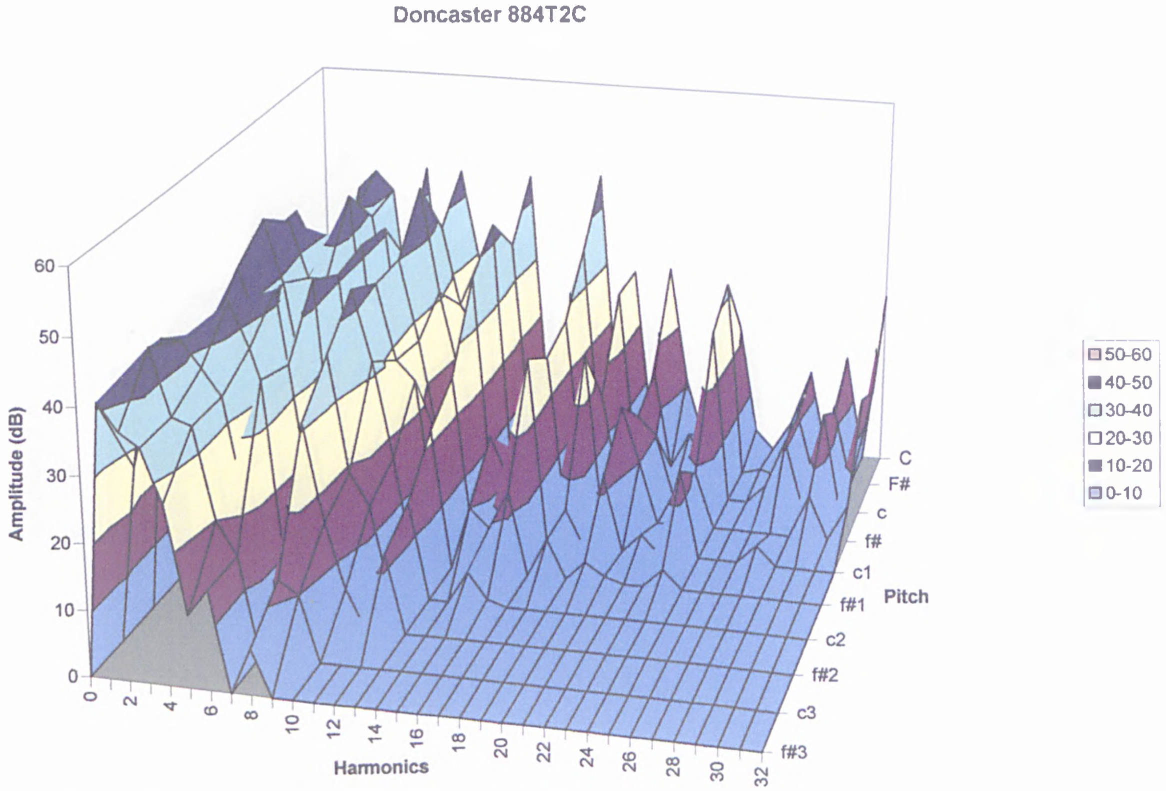


Figure 5.3 – Acoustic signature for the Doncaster Parish Church ensemble

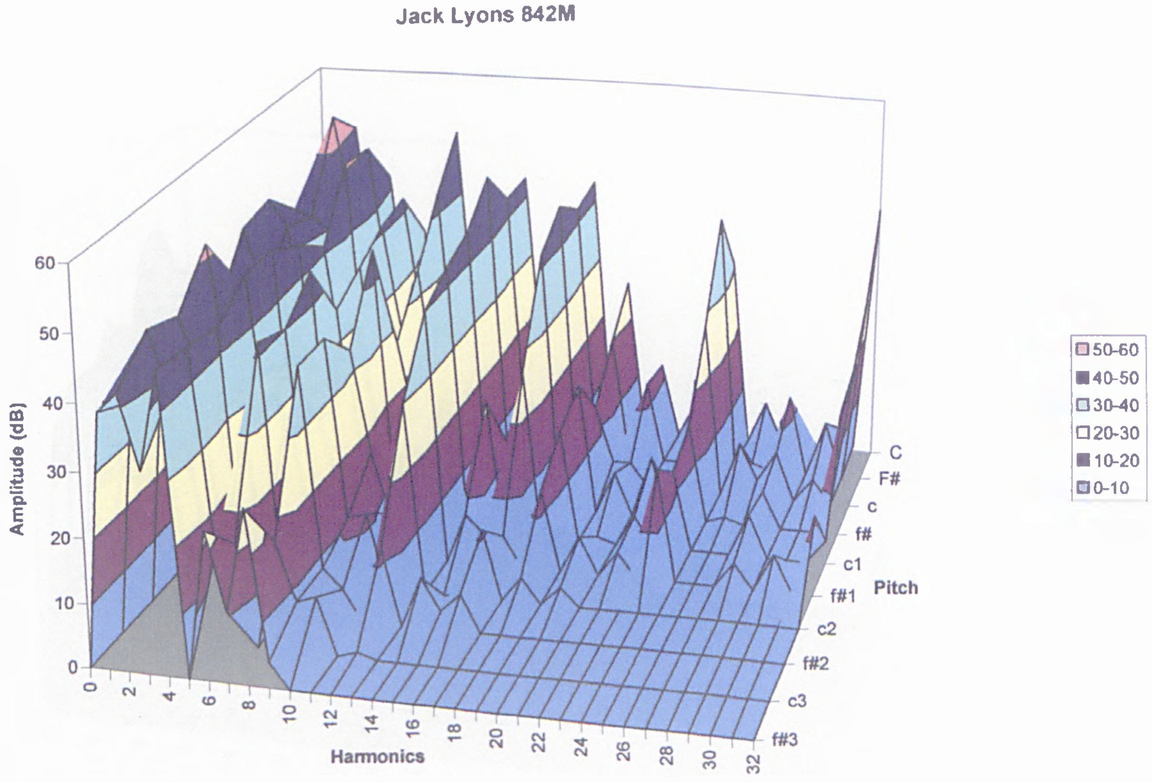


Figure 5.4 – Acoustic signature for the Jack Lyons ensemble

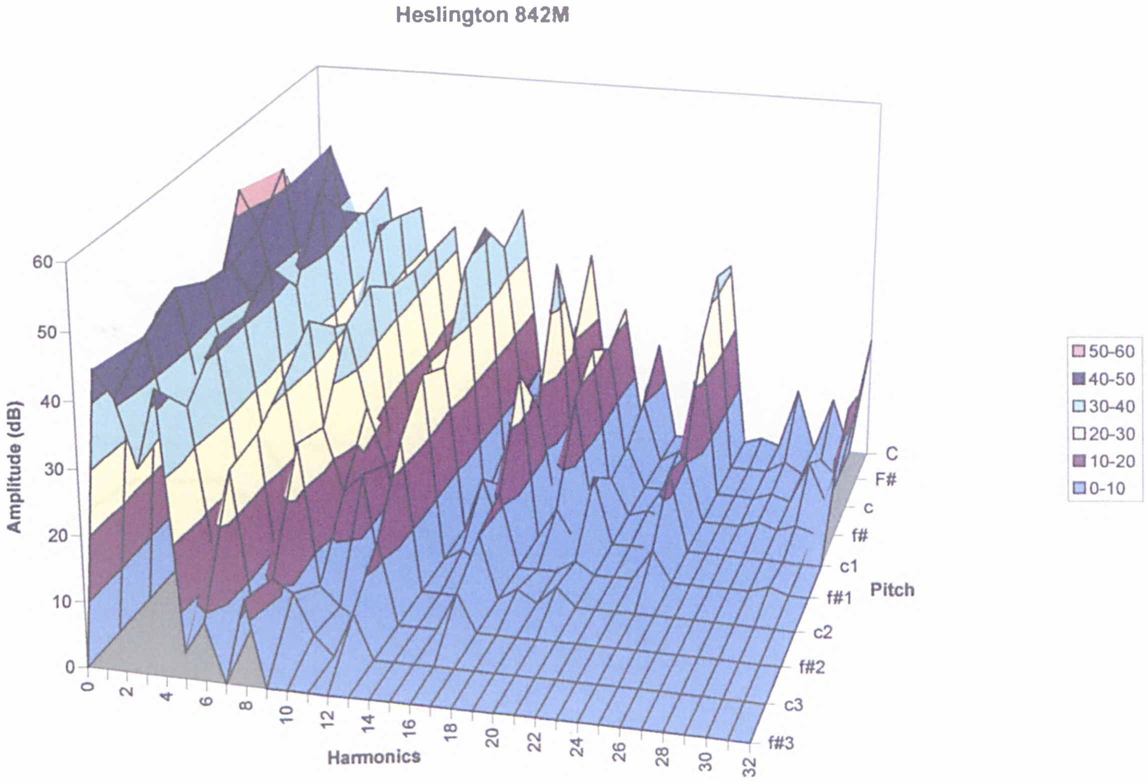


Figure 5.5 – Acoustic signature for the Heslington ensemble

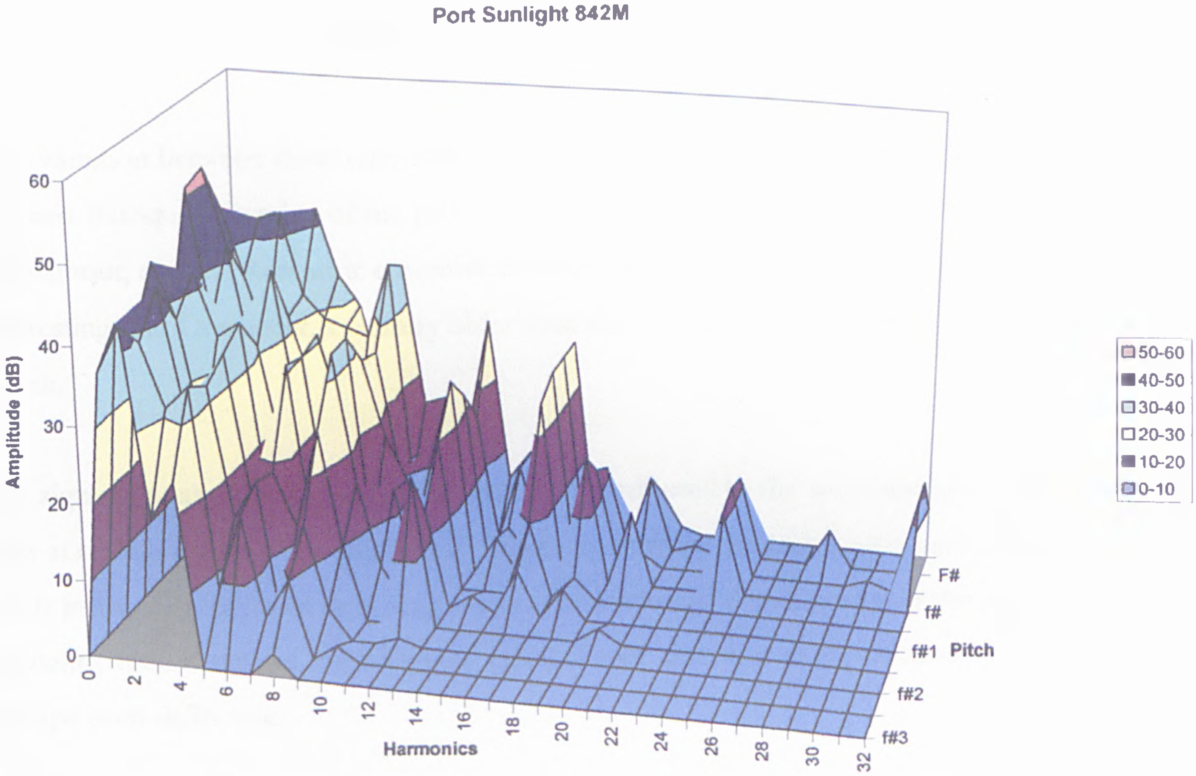


Figure 5.6 – Acoustic signature for the Port Sunlight ensemble

All of the comparative graphs in this section display all six ensembles used in the general and specialised listener tests. Full numerical data for the analyses below can be found in appendix J. Figure 5.7 shows the normalised spectral centroids for the six ensembles.

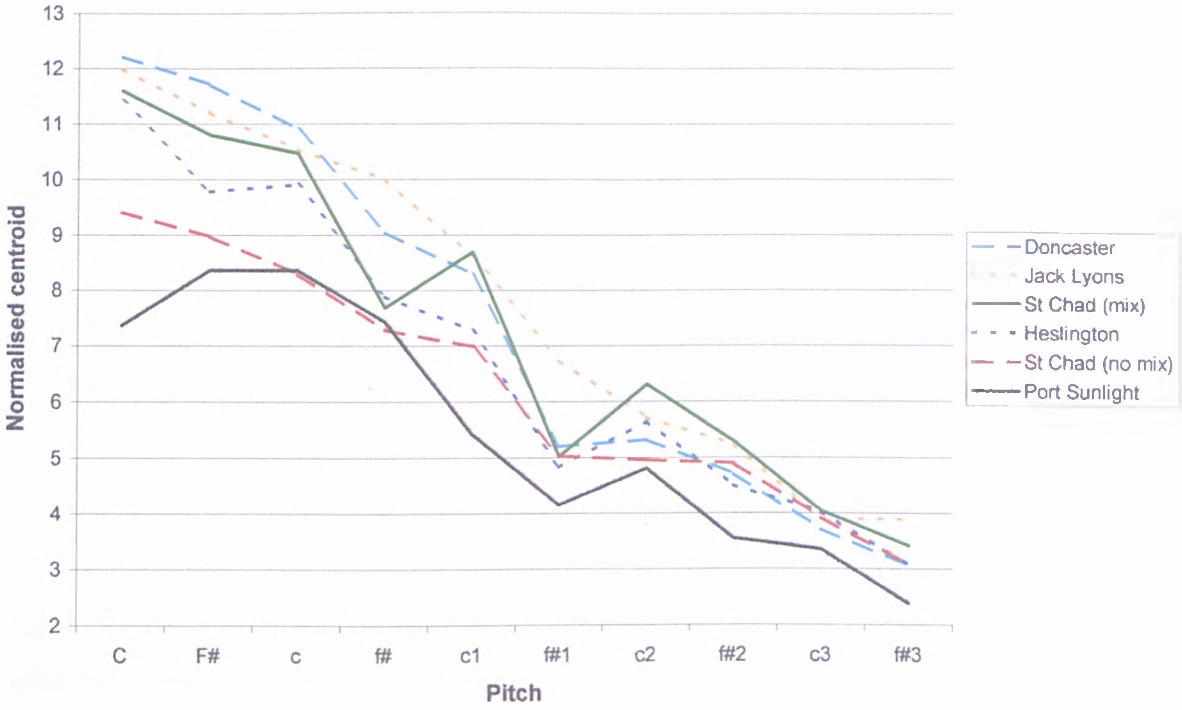


Figure 5.7 – Normalised spectral centroids

The variation between these ensembles is most interesting. Perhaps not surprisingly, the neo-Baroque ensemble of the Jack Lyons has one of the highest spectral centroids throughout, and the Romantic ensemble of Port Sunlight has one of the lowest. It is interesting that Doncaster, a century older than the Jack Lyons, is overall its closest match.

The zigzagging around the centre can mostly be attributed to the way mixtures break back at certain points. The fact that all organs apart from the Jack Lyons have a dip at f#1 is interesting, as it not only suggests uneven harmonic development in the most frequently used octave of the keyboard, but that such unevenness is common and perhaps even deliberate.

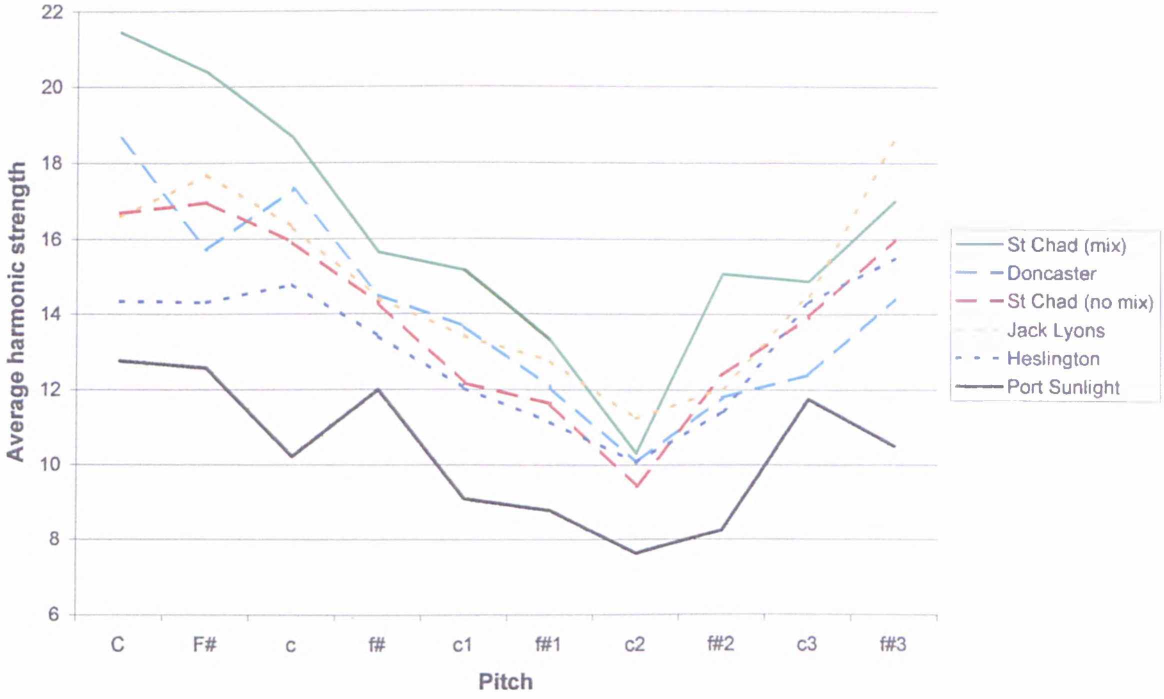


Figure 5.8 – Average harmonic strength

Figure 5.8 shows that the ensemble with the highest centroid is not necessarily the one with the strongest harmonics. St Chad is significantly the strongest, with Port Sunlight clearly the weakest. All have a dip at c2, which is possibly due to the exclusion of harmonics above the Nyquist frequency for the top three pitches.

There is also significant difference in the smoothness or consistency over the keyboard range between the six ensembles. Heslington is noticeably smoother than most, whereas Doncaster is relatively uneven. Port Sunlight is uneven at the extremes, but has less of a dip in the centre, resulting in a mathematical evenness score of less than St Chad's. Both of the St Chad's ensembles do not change direction frequently, but receive a high unevenness score because of the steepness of their slopes. Evenness scores (smoothness across the keyboard range) can be found along with the data from which figure 5.8 is derived in appendix J.

The reader's attention is particularly drawn to the fact that the y-axis scale on this and other graphs in this section does not start at zero. This is to permit the best presentation of the significant sections of data.

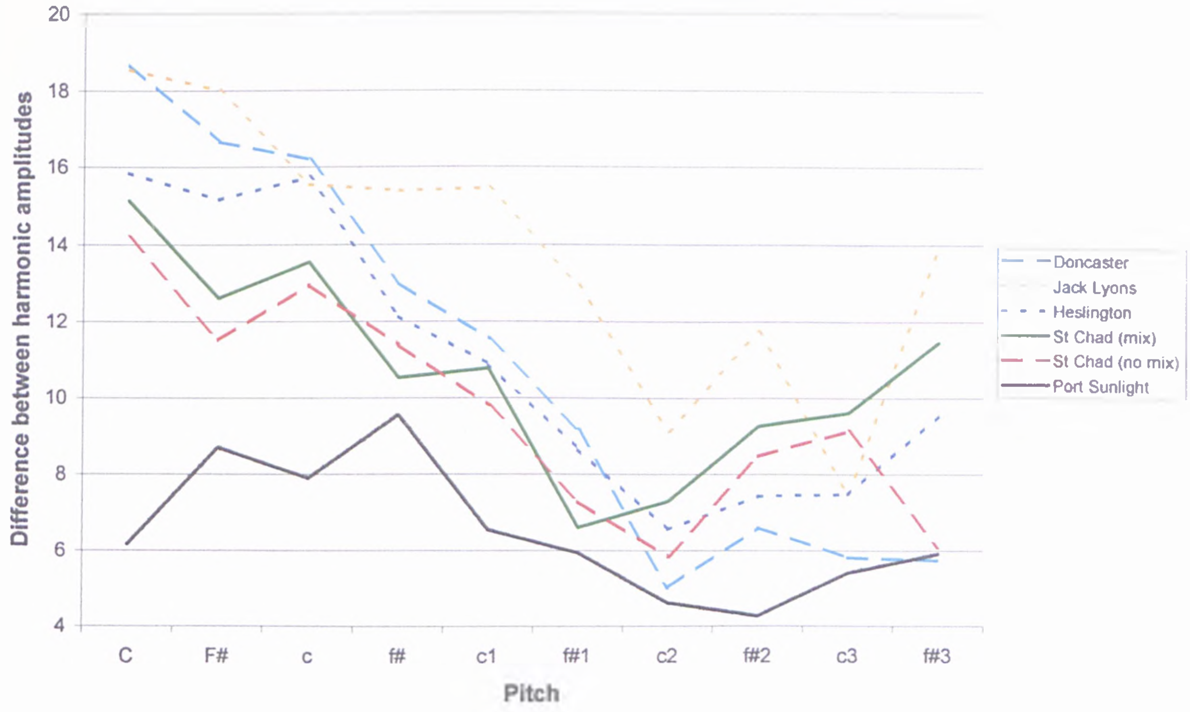


Figure 5.9 – Spectral smoothness

Figure 5.9 displays the way spectral smoothness (the average absolute difference between one harmonic and the next) changes over the keyboard range. A low score represents a smoother transition. Overall, the Jack Lyons is the least smooth ensemble, but things are less clear for all ensembles among the highest pitches.

Figures 5.10 and 5.11 respectively display spectral slope (the difference in average amplitude of low and high frequencies) and inter-quartile spectral slope for all six ensembles. Similarities and differences between the two different measures are clearly visible, and it remains to be seen which will prove the most useful. Both graphs have their amplitude difference scale in decibels, and a lower score represents a shallower spectral slope.

All of the graphs in figures 5.7 to 5.11 have noticeable differences at the highest frequencies. This is probably due to the fact that the Nyquist frequency is sufficiently low to omit, for example, the fourth harmonic of the Fifteenth 2' component in each ensemble for f#3. Additionally, the upper octave of the keyboard is less used in music (not at all in the musical examples used in these tests) and harder to regulate than the other octaves due to the small size of the pipes. It is therefore proposed to omit the upper two samples from the mathematical calculations of spectral properties to avoid any artefacts that these factors might introduce.

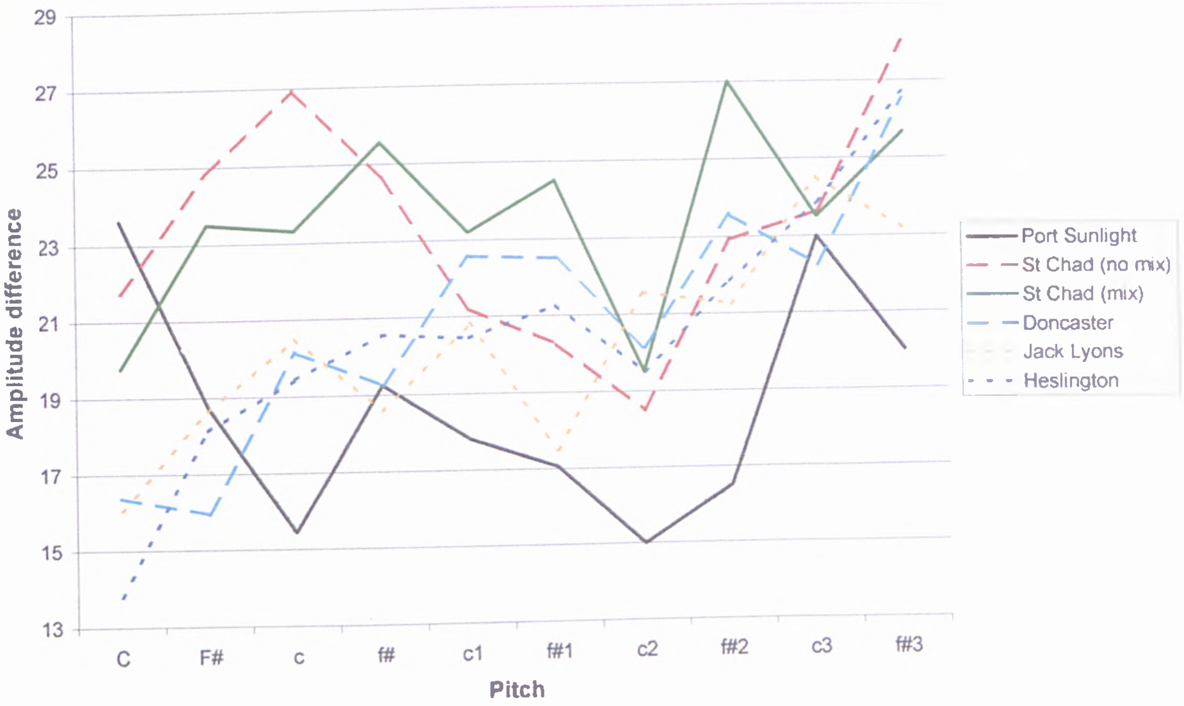


Figure 5.10 – Spectral slope

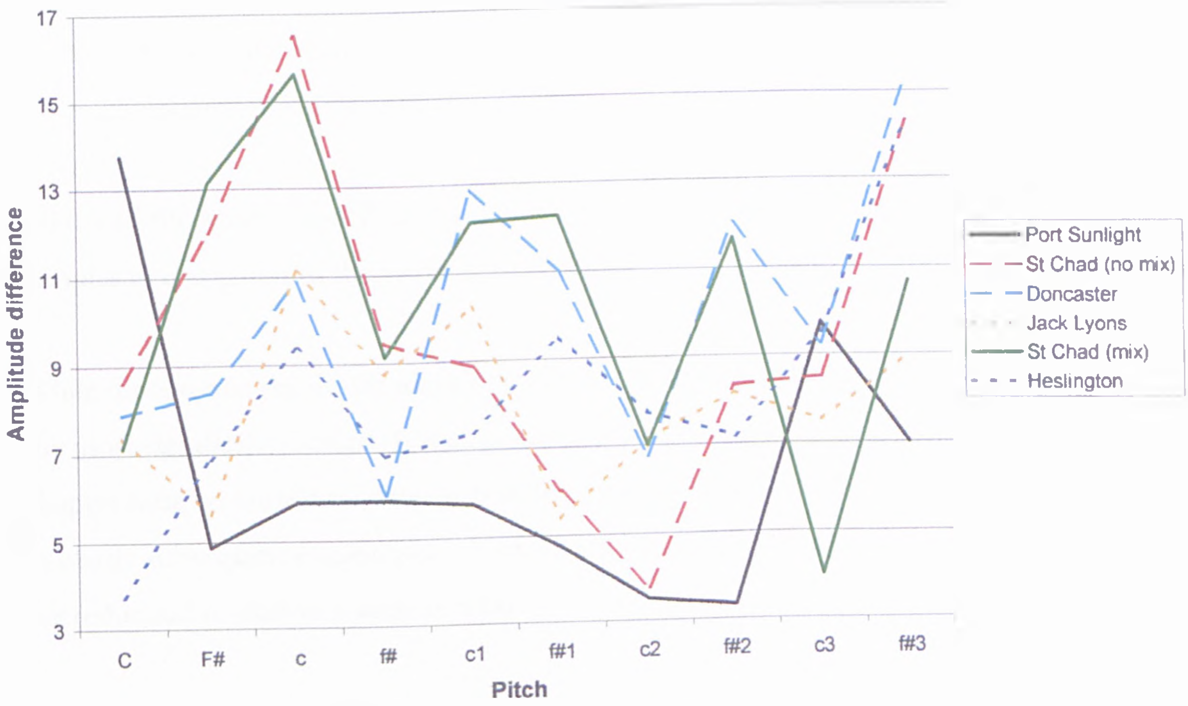


Figure 5.11 – Inter-quartile spectral slope

5.7.3 Reduction of data sets

Each ensemble has a number of readily isolated steady-state spectral features described in section 5.7.1. In summary, these features are:

- Spectral centroid (normalised units of pitch)
- Centroid fall-off (normalised units of pitch)
- Centroid consistency (normalised units of pitch)
- Average harmonic strength (dB)
- Strength fall-off (dB)
- Strength consistency (dB)
- Spectral smoothness (dB)
- Standard deviation of spectral smoothness
- Average spectral slope (dB)
- Standard deviation of spectral slope
- Average inter-quartile spectral slope (dB)
- Standard deviation of inter-quartile spectral slope

All use of the term "fall-off" refers to a reduction of that quantity in the higher note samples as compared to the lower note samples.

While these terms are readily identifiable from the spectral data, it is unclear which are the most significant factors. Just as potentially correlated words were removed in chapter four, so correlated data derived from acoustic analyses should also be removed to clarify subsequent comparison. Various mathematical and statistical techniques exist for reduction of data sets such as these.

Conventional multi-dimensional scaling is inappropriate as it both seeks to describe a particular timbre space, for which we have an inadequate number of samples, and to do so by a minimal number of dimensions. It would be inappropriate to suggest that the features above represent all possible spectral correlates, particularly as they ignore the transient portion of the sound, and thus MDS techniques are inappropriate.

More appropriate would be principal component analysis, which identifies the degree to which a number of factors contribute to theoretical principal components in a decreasing order of importance. PCA does not attempt to describe the entire timbre space with a minimal number of dimensions in the way of MDS, but assumes that some components are perceptually less important and can be discarded. More details of PCA can be found in Manly (1994).

It should be noted that Creasey (1998, p48) disputes the wholesale discarding of less important components in this manner as flawed. Frequently the first component can be responsible for the majority of difference between the samples. However, these theoretical components do not necessarily correlate to the input variables, and thus do not directly help identify significant variables. PCA is a procedure that works best for more general statistical data, and its automatic application in the area of psychoacoustics is questionable given the difficulty in isolating acoustic factors and the complex interaction of perception.

Some parts of PCA are still useful here as many of the potential spectral components identified could be correlated, and thus if all were used it would be difficult to isolate a particular variable to alter in the synthesis process. The values of all potential spectral components are presented in table 5.16.

Potential spectral components	Port Sunlight	Jack Lyons	Heslington	Doncaster	St Chad (mixture)	St Chad (no mixt.)
Spectral centroid (NFU)	6.184	8.751	7.654	8.424	8.233	6.979
Centroid fall-off (NFU)	3.395	4.359	4.194	5.091	3.813	3.030
Centroid consistency	1.019	0.965	1.264	1.104	1.561	0.646
Average harmonic strength (dB)	10.160	14.303	12.685	14.216	16.249	13.684
Strength fall-off (dB)	3.445	3.895	3.050	4.648	5.568	4.563
Strength consistency	1.329	1.199	0.940	1.926	2.271	1.536
Spectral smoothness (dB)	6.710	14.594	11.546	12.105	10.718	10.183
Standard deviation of spectral smoothness (dB)	1.883	3.168	3.783	4.943	2.973	2.859
Average spectral slope (dB)	17.930	19.359	19.370	20.059	23.289	22.628
Standard deviation of spectral slope	2.739	1.976	2.525	2.823	2.600	2.754
Average inter-quartile spectral slope (dB)	5.965	7.978	7.336	9.471	11.009	9.193
Standard deviation of inter-quartile spectral slope	3.302	2.038	1.786	2.550	3.017	3.823

Table 5.16 – Values of potential spectral components

The aim of factor analysis is to eliminate highly correlated variables. The Pearson correlation coefficients between all twelve potential spectral correlates are given in table 5.17 below. Coefficients significant at the 5% level for a sample size of six are marked in bold.

	Centroid	Fall-off	Consist.	Strength	Fall-off	Consist.	Smooth	Std Dev.	Spec Slope	Std Dev.	IQSS	Std Dev.
Centroid	**1**	**0.749**	0.368	**0.790**	0.375	0.277	**0.912**	0.666	0.235	-0.537	0.566	-0.615
Fall-off	**0.749**	**1**	0.346	0.302	0.001	0.098	0.635	**0.845**	-0.281	-0.204	0.159	**-0.733**
Consist.	0.368	0.346	**1**	0.392	0.272	0.405	0.049	0.170	0.154	-0.020	0.342	-0.397
Strength	**0.790**	0.302	0.392	**1**	**0.794**	0.645	0.633	0.427	**0.775**	-0.225	**0.929**	-0.107
Fall-off	0.375	0.001	0.272	**0.794**	**1**	**0.946**	0.112	0.145	**0.837**	0.230	**0.926**	0.444
Consist.	0.277	0.098	0.405	0.645	**0.946**	**1**	-0.067	0.179	0.679	0.410	**0.830**	0.430
Smooth	**0.912**	0.635	0.049	0.633	0.112	-0.067	**1**	0.612	0.122	-0.687	0.355	-0.660
Std Dev.	0.666	**0.845**	0.170	0.427	0.145	0.179	0.612	**1**	0.062	0.089	0.392	-0.514
Spec Slope	0.235	-0.281	0.154	**0.775**	**0.837**	0.679	0.122	0.062	**1**	0.204	**0.891**	0.446
Std Dev.	-0.537	-0.204	-0.020	-0.225	0.230	0.410	-0.687	0.089	0.204	**1**	0.139	0.594
IQSS	0.566	0.159	0.342	**0.929**	**0.926**	**0.830**	0.355	0.392	**0.891**	0.139	**1**	0.191
Std Dev.	-0.615	**-0.733**	-0.397	-0.107	0.444	0.430	-0.660	-0.514	0.446	0.594	0.191	**1**

Table 5.17 – Correlation between potential spectral components

A large number of these variables have significant correlation with other values, including some that are not immediately obvious. It must be emphasised that the correlations between variables in table 5.17 are only applicable to the set of samples presented here. Instead of PCA, it is possible to reduce the data set while remaining within measured spectral qualities by choosing one variable at a time and eliminating those variables it correlates with. Thus features mentioned in previous tests can be prioritised to allow confirmation or rejection of earlier theories. A full list of qualities from the previous research mentioned in section 5.1 follows:

- Spectral centroid
- Spectral width
- Number of harmonics
- Harmonic density
- Spectral slope
- Amplitude of second harmonic
- Presence or lack of lower harmonics
- Having both odd and even harmonics

As each sample's spectrum has a significant fundamental component and similar tendencies to fall-off at the upper frequencies, spectral width and harmonic density can be determined by comparing the centroid and average harmonic strength. As these are significantly correlated, the resulting measure would also be highly correlated with both.

It would be impossible therefore to discern which of the two theories of "brightness" mentioned in section 5.1 is generally applicable, as both would seem applicable.

As the organ's steady-state spectrum is harmonic in nature, the number of harmonics is also effectively the same as the spectral width. It is entirely possible that a more varied set of samples would produce less correlation between these qualities, but in the context of this test it would be inappropriate to use more than one of these. As the spectral centroid is the most widely recognised and easily determined of these qualities, it is proposed to prioritise its use over those variables it correlates highly with.

The comments on lower harmonics are interesting, as all samples have broadly similar lower harmonic strength. It should be possible at the synthesis stage to produce more variety of levels in the lower harmonics. The theory of "fullness" that suggested it relied upon samples having both odd and even harmonics is also difficult to apply to this particular sample set. All samples possess odd harmonics at lower frequencies, but the higher frequencies tend to be only even harmonics. It must also be remembered that the term "full" was rejected in section 5.6.3 as lacking in common understanding, so verification of its associated previous theory is not possible.

It is possible to add some mathematical reasoning to this process by looking at the overall degree to which each word correlates with all others. Table 5.18 presents the sum of the squared correlation between each variable and all others. This is inspired by the relationship between factor loadings and matrix eigenvalues in factor analysis, but actual factor analysis is inappropriate as the context is entirely different. It provides an absolute comparison emphasising those variables with high correlation.

Centroid	Fall-off	Consist.	Strength	Fall-off	Consist.	Smooth	Std. Dev.	Spec. Slope	Std. Dev.	IQSS	Std. Dev.
4.855	3.582	1.977	5.024	4.582	4.101	4.078	3.226	3.973	2.496	5.001	3.755

Table 5.18 – Total squared correlation of potential spectral components

A lower value in table 5.18 indicates a variable more independent of the other variables. Variables having a high total squared correlation are not necessarily poor choices, but variables with a lower total are more likely on average to be useful. However the data from tables 5.17 and 5.18 must also be compared with the values in figures 5.7 to 5.11 to allow informed selection of variables.

From both those tables, consistency of the spectral centroid across the keyboard range would appear to be the least correlated variable. Fall-off of the centroid correlates significantly with three other terms, so will not be used. Measurements of strength all have significant correlation with other variables and within their own group, so will also not be used. The measurements of spectral slope have correlation only with variables now rejected, and of the two, conventional spectral slope has less correlation than inter-quartile spectral slope and will thus be used instead. The standard deviation of spectral slope, a measure of how consistent the spectral slope is over the keyboard range, is also uncorrelated with any other variable and will therefore be selected.

The measure of spectral smoothness correlates highly with the spectral centroid and thus is unsuitable for further testing. The standard deviation of smoothness, which measures how consistent each ensemble is in its smoothness over the keyboard range, correlates highly but not significantly with the centroid value, and significantly with centroid fall-off, rejected due to its correlation with the centroid. It would also be counter-intuitive to include a variable essentially dependant on one previously rejected.

This leaves four potential spectral features that are not correlated and are readily identified from spectral analyses. These four are:

- Spectral centroid
- Consistency of spectral centroid over the keyboard range
- Spectral slope
- Consistency of spectral slope over the keyboard range

It is interesting at this point to consider the only spectral theory not yet covered, namely that of clarity relating to a strong second harmonic. The Pearson correlation coefficients for all the remaining variables are presented along with those for the average strength of the second harmonic and its fall-off from low to high keyboard range. Significant correlation at the 5% level is indicated in bold.

	Centroid	Consistency	Spectral slope	Std. Dev.	2^{nd} harm.	Fall-off
Centroid	**1**	0.368	0.235	-0.537	0.164	-0.096
Consistency	0.368	**1**	0.154	-0.020	-0.087	-0.080
Spectral slope	0.235	0.154	**1**	0.204	**0.761**	0.051
Std. Dev.	-0.537	-0.020	0.204	**1**	-0.303	-0.600
2^{nd} harmonic	0.164	-0.087	**0.761**	-0.303	**1**	0.465
Fall-off	-0.096	-0.080	0.051	-0.600	0.465	**1**

Table 5.19 – Correlation between remaining spectral and second harmonic variables

The significant correlation indicates that ensembles with a greater spectral slope also tend to have a stronger second harmonic component. This is perhaps intuitive. The fall-off of the second harmonic tends towards negative correlation with the standard deviation of spectral slope, but not to a significant degree. This suggests that those ensembles having less consistency in their spectral slope over the keyboard range are likely also to be those with a second harmonic stronger in the treble than the bass. The correlation of the second harmonic strength with spectral slope means that it would be inappropriate to include it as a separate measure. Again, it would be unintuitive to include its related fall-off despite it not reaching significant levels of correlation with any remaining variable.

This section demonstrates that many of the potential spectral correlates identified in previous studies are highly correlated in the timbre space of the pipe organ principal ensemble. It will be difficult therefore to say with certainty that one particular theory has been supported when that theory has other correlated variables. However, as the goal of this section is to provide guidance as to the synthesis of examples to demonstrate the common understanding suggested by experimental results thus far, such multiplicity is not necessarily confusing.

Indeed, the correlation shown between variable suggests that it may be difficult to implement theories in isolation. Creasey (1998, p45) agrees that it is impossible to have completely unbiased parameters in this kind of study, and questions the usefulness of attempting to isolate such parameters (p44), although in this thesis its usefulness and limits are clear. The correlation between spectral slope and the strength of the second harmonic suggests that alteration of the second harmonic strength may be a logical way of adjusting spectral slope.

5.7.4 Correlation between remaining adjectives and spectral features

The simplest way to determine a link between commonly understood adjectives and potential spectral features is to examine the correlation between the listener ratings and the measured spectral features. The correlation between the remaining timbral adjectives has been given in table 5.14, and the correlation of remaining spectral features has been given in table 5.19. Table 5.20 presents the correlation between timbral adjectives and measured spectral features. Correlation significant at the 5% level is indicated in bold.

Adjective	Spectral centroid	Consistency of centroid	Spectral slope	Standard deviation of slope
Thin	0.775	-0.382	-0.097	**-0.969**
Flutey	**-0.952**	0.044	-0.275	**0.913**
Warm	-0.747	0.431	0.139	**0.976**

Table 5.20 – Correlation between remaining spectral variables and timbral adjectives

Thin and warm both tend towards significance, so at this point some of the potential spectral variables previously rejected on grounds of correlation with the measure of spectral centroid can be brought back in to see if any of them correlate better with these adjectives. Table 5.21 shows those variables and their correlation.

Adjective	Centroid fall-off	Average strength	Strength fall-off	Strength consistency	Smoothness	Standard deviation of smoothness
Thin	**0.929**	0.352	-0.178	-0.423	**0.941**	0.617
Flutey	**-0.930**	-0.670	-0.166	0.091	**-0.996**	-0.700
Warm	**-0.907**	-0.314	0.201	0.442	**-0.923**	-0.574

Table 5.21 – Correlation between additional spectral variables and timbral adjectives

5.7.5 Development of spectral correlation theories

To avoid confusion, the variables that correlate highly with the remaining adjectives will be summarised here. All the adjectives work on a simple scale such that a higher value for a particular word indicates its greater perceived presence.

- Spectral centroid: a higher value in normalised frequency units indicates a higher average to the entire spectrum.
- Standard deviation of spectral slope: a higher value indicates a greater inconsistency between the spectral slope over the keyboard range.
- Centroid fall-off: all organs have a decrease in normalised centroid frequency at higher notes on the keyboard, and a higher value indicates when this effect is greater.
- Smoothness: this measures the degree to which each harmonic is different in level from the next, and averages this over the keyboard range. A greater smoothness rating indicates less spectral smoothness, which may seem counter-intuitive. To name it "spectral" roughness is however inviting confusion with perceived roughness, a defined psychoacoustic measure.

"Flutey" has a high negative correlation with the spectral centroid, meaning that it is associated with a reduction in centroid frequency. Both "flutey" and "warm" have positive correlation with the standard deviation of spectral slope, suggesting association with those ensembles featuring greater variation in slope over the keyboard range.

"Thin", on the other hand, is negatively correlated with this factor, suggesting an association with those ensembles maintaining a consistent spectral slope. However, looking at figure 5.10 suggests that in particular Heslington and Port Sunlight might have their standard deviations affected by the results for the lowest pitch, as their general tendency in the mid-keyboard range is for less variation than in the extremes. This casts some doubt on the strength of correlation between the consistency of spectral slope and the adjectives in question, but does not rule out an important connection.

All three adjectives have high correlation with the fall-off in frequency of the spectral centroid over the keyboard range. For "thin" this is positive correlation, meaning that a greater fall-off accompanies use of the word "thin", and for the other two words the correlation is negative. Smoothness is similar: "thin" accompanies the least smooth ensembles, and the other two adjectives accompany smooth ensembles. In the case of "flutey", this correlation is very strong, remaining significant at even the 0.5% level.

The other categories do not have any significant correlation with adjective occurrence, although there are some trends. "Thin" tends towards correlation with a higher spectral centroid, and "warm" tends towards correlation with a lower spectral centroid. These and other tendencies would be statistically significant if their values were maintained when a larger number of organs were compared.

"Flutey" tends towards correlation with lower average harmonic strength and a more consistently smooth ensemble, whereas "thin" tends towards correlation with a less consistently smooth ensemble. However as these relationships are all below the threshold of statistical significance, not too much can be drawn from them.

It is interesting to note that despite the correlation between average harmonic strength and the spectral centroid, this is not matched by similar results for the adjectives in both of those categories. It may be worthwhile to consider some of the frequency-related suggestions from previous theories for each adjective in light of this, despite the apparent direct correlation between second harmonic strength and spectral slope.

Do these correlations match what has been predicted in the theories presented in section 5.1? The three adjectives will be examined individually before the results for "bright" and "clear" from the general listening test are examined.

"Thin" was said in previous research to be associated with a lack of lower frequency components. Of the measures developed in this research, that would match most closely with an increase in spectral centroid. In practice there is a tendency towards a positive correlation with increased spectral centroid frequency. Positive correlation also exists with a consistent spectral slope, significant decrease in centroid towards the treble, and those ensembles that are less acoustically smooth (those with prominent harmonics). It will be interesting to compare this with a measure of harmonic strength for the lower harmonics to see if there is a negative correlation.

"Flutey" had no previous theories, as this would seem to be a word with a particular meaning in the context of the pipe organ. However, it does exhibit common understanding, and correlates with a lower spectral centroid, greater variation in spectral slope across the keyboard range, less centroid fall-off from low to high notes, and strongest of all with a smooth spectrum, with a minimum of prominent harmonics.

Warmth was theorised by previous studies to be associated with a decrease in the number of harmonics, which in this context equates to a decrease in spectral centroid, and an increase in energy at lower frequencies, particularly first three. When the lower harmonics are studied, it would make sense to restrict these to the first three both to examine that particular theory and as more general measures of harmonic strength have not demonstrated any significant correlation with any of the adjectives.

Warmth was also associated in previous research with compressed harmonic density and decreased spectral slope. Harmonic density in this context can be determined from the centroid frequency and average harmonic strength – a less harmonically dense spectrum will have the same centroid but less average strength. However in the examples analysed both qualities are correlated. Strength also has no significant correlation with the adjectives, so it is difficult to comment on harmonic density within the constraints of this sample set.

As some of the previous theories suggest links between lower harmonics and certain adjectives, a final set of correlation between the three adjectives and three lower harmonic strength measures is presented in table 5.22.

Adjective	Strength of first harmonic	Fall-off of first harmonic	Strength of second harmonic	Fall-off of second harmonic	Av. strength of first three harmonics	Fall-off of first three harmonics
Thin	-0.568	-0.749	0.455	0.242	0.038	-0.789
Flutey	0.657	0.454	-0.749	-0.267	0.645	0.645
Warm	0.582	0.777	-0.415	-0.275	0.776	0.766

Table 5.22 – Correlation between lower harmonic properties and timbral adjectives

None of the results in table 5.22 reach statistical significance, although there are some interesting trends. "Thin" sounds tend towards a weaker fundamental and stronger second harmonic, which is interesting in light of the other correlation above in table 5.21 and previous theories. "Flutey" sounds tend towards strong first harmonics, but away from prominent second harmonics. Warm sounds tend towards a strong fundamental, but as with all these results they are not significant, unlike some of the correlation data previously presented.

It is impossible to present correlation data for the other two words tested in the large-scale general listener test, as the ensembles used there were not considered exhaustively. Data on the average and standard deviations for all subsets of the listeners has already been presented and discussed in section 5.4, but standard deviations were on the whole greater than for the specialised listeners. In most cases all subsets broadly follow the general trend, so it is valid to consider the overall average answer in most cases. Table 5.23 presents for each ensemble used in the general listening test the spectral variables either mentioned in previous theories or which have appeared significant for other adjectives.

Organ	Average lower harmonic strength	Average strength of second harmonic	Spectral centroid	Centroid fall-off	Spectral smoothness	Spectral slope	Standard deviation of spectral slope
Doncaster	38.667	39.5	8.424	5.091	12.105	20.059	2.823
Jack Lyons	39.25	46	8.751	4.359	14.594	19.359	1.976
St Chad's (no mixture)	43.458	49.375	6.979	3.030	10.183	22.628	2.754
St Chad's (mixture)	42.917	47.875	8.233	3.813	10.718	23.289	2.600

Table 5.23 – Spectral variables for large-scale test audio examples

In the case of "bright", the pair of samples from St Chad's appears to confirm the well-established theory that a higher centroid is perceived as brighter (by 16.4% of the possible maximum), even though the mixture added was more complex harmonically. It is interesting to note that subject preference was biased 18% towards the example without the mixture. The average result of the other pair of samples, with a 22.2% bias towards Doncaster, would appear to confuse this theory, until the results from those who indicated some experience with pipe organs are taken into account.

That subgroup had a 12.4% bias in the opposite direction, suggesting that experience of pipe organs plays an important part in the perception of brightness. For the more general subjects, other factors, perhaps including the very different acoustic environments of the organs, clearly distracted from the relatively fine judgement of which sample had the higher centroid. As the focus of this thesis is on those listeners with experience of pipe organs, it is inappropriate to conduct further tests to examine the difference between them and general listeners, but the difference is real, intriguing and worthy of further study outside this thesis.

The addition of the mixture made little difference to perceived clarity in the case of the two St Chad's samples, but the comparison of the Jack Lyons and Doncaster ensembles gave the latter a bias of 22.2%. Those listeners with pipe organ experience also concurred with this opinion, although on balance preferring the Doncaster ensemble to a lesser extent than the Jack Lyons.

Two contrasting theories introduced in section 5.1 related perceived clarity to a decreased number of harmonics and an increase in higher harmonics. The spectral centroids of the four ensembles seem not to support any of those theories, as there is a much greater difference in the case of St Chad's two ensembles. Ethington and Punch (1994) suggest decreased spectral slope is also a cause, but the slight difference there is between the Doncaster and Jack Lyons ensembles on that scale is the opposite way round. The suggestion of Jeans (1961) that it relates to a stronger second harmonic is also at odds with the values in table 5.23. Clarity, at least in this context, seems unrelated to either the spectral centroid or any of the previously suggested theories.

What none of the steady-state analyses can take fully into account is the acoustical environment in which the organ is situated and the effect of that on listener perception. It might be expected, for example, that a more reverberant acoustic would be perceived as less clear, but the comparison of the Doncaster and Jack Lyons organs suggests the reverse. It would be interesting to see if this was the case with a truly identical ensemble in different acoustic environments, and the synthesis portion of this study will permit answering of that question.

It may be that reverberation is also the significant difference between the adjectives "warm" and "flutey". Examination of the raw results suggests that that majority of the difference between those terms comes where the mixture ensemble of St Chad's is pitched against those of Heslington and Jack Lyons. Although a direct comparison of Port Sunlight with St Chad's results in a tie between those terms (both biased towards Port Sunlight), the other results suggest that a more reverberant ensemble might be considered "warmer" but not "flutier". Again, the synthesis stage should allow examination of this possibility.

5.8 Summary of theories developed

Section 5.7 includes a large amount of data, which it is useful to summarise by adjective here. "Balanced" and "full" exhibited no common understandings in the context of this test, and have therefore been excluded from further study.

- **Thin**: Ensembles described as "thin" tended towards a higher spectral centroid, and correlated with a greater fall-off over the keyboard range. There was a high correlation with an uneven ensemble (one with prominent harmonics), but also a possible correlation with a more consistent spectral slope. It is unclear at this stage whether it is the lack of lower frequencies (as suggested by previous research) or the presence of prominent higher harmonics (as suggested here) that contributes to a perception of "thin".

- **Flutey**: Ensembles described as "flutey" had significant correlation with a lower spectral centroid and less centroid fall-off over the keyboard range. There was very high correlation with a smooth ensemble (one without prominent harmonics) and possible correlation with a more inconsistent spectral slope from note to note.

- **Warm**: Ensembles described as "warm" tended towards correlation with a lower spectral centroid and correlated with centroid fall-off and increased smoothness, but to a lesser extent than "flutey". They also correlated to a greater extent with an inconsistent spectral slope. There was a tendency towards correlation with strong lower harmonics, but little to explain the differences between "warm" and "flutey". The amount of reverberation was also suggested as a possible trigger.

- **Bright**: Overall the previous theories linking a perception of brightness and an increased spectral centroid were maintained. Due to the nature of the samples under consideration, it has not been possible to distinguish between the theories of spectral centroid and spectral width. It is not easy to conceive a test within the timbral constraints of this study that could do so.

- **Clear**: None of the previous theories appeared to be supported by the evidence of the large group test, although common understanding was demonstrated. It will be interesting to examine whether any theories, either new or old, will be suggested by the results of the synthesis experiment.

All of these words are worthy of continued study despite possessing different levels of certainty as to their spectral correlates. More conclusions can be drawn once they have been studied in the synthesis experiment described in chapter six.

6. Verification of theories

This chapter attempts to verify the theories discussed and developed in section 5.7 by synthesising a small number of ensembles and using them in a further listening test. The process of synthesis to prove theories developed by analysis has been introduced as a concept by Risset and Wessel (1999, p141). It permits the testing of theories by the theoretical removal of all variables apart from those being tested, and the opportunity to isolate factors. There is then the possibility of using standard techniques such as PCA or MDS on the results of listener tests, the latter being practically impossible with conventional complex samples.

6.1 Experiments to verify proposed theories

6.1.1 Development of methodology

The initial plan described in section 1.5 and developed throughout the thesis was to use the Bradford synthesis system to create a number of ensembles with different characteristics designed to test the theories developed in section 5.7. Two arguments against this methodology had become apparent.

Firstly, there were good reasons for keeping the synthesis within the realms of what was possible on a single pipe organ. The results would therefore be directly applicable to an organist playing a reasonably-sized instrument. Multiple ensembles would inevitably result in an increase in the number of factors changed between each synthesised ensemble, and thus make it harder to isolate individual relationships between adjectives and spectral features.

Secondly, as mentioned in section 3.2, the experimental organ had become unreliable in operation, particularly when synthesising ensembles. It was still possible to use the organists' long established method of synthesis by addition or subtraction of stops from the selection available. While lacking the precise control of the initially proposed methodology, this method could still implement the theories developed.

The question of the effect of reverberation could also be addressed by playing the same ensemble through different settings on a high quality reverberation device. The unreliability of the MIDI input device meant that each example had to be played manually, but this was also the case with the real-world examples used throughout chapter five.

The artificial examples were accordingly created using this procedure, and the choice of stops and other details are described in section 6.1.2. It was then planned to record a set of short note-by-note extracts to create acoustic signatures in the same way as for the ensembles analysed in section 5.6. Unfortunately, the increasingly unreliable experimental equipment prevented this. It proved possible with repeated attempts to create and record the short musical extracts, but immediately thereafter the organ became so unreliable that even recordings of single-note samples for analysis became impossible, despite repeated attempts over the author's remaining experimental time.

The main graphs in this chapter, therefore, resort to the harmonic amplitude method used in section 3.2. While they are directly comparable with each other, as they again take no account of the relative volumes of each stop, they cannot be regarded as accurate harmonic analyses. It is likely that the effect of the mixture in particular is exaggerated in these. As an alternative attempt to analyse the actual output, a short section of the final chord of each musical extract has been analysed and subjected to those analyses still possible, but the results from those cannot be directly compared with previous analyses as they are not from single note samples. They do however give a means of comparing the four synthesised ensembles with each other.

6.1.2 Experimental procedure

Although the synthesis procedure outlined in section 6.1.1 is unconventional, as a means of theory validation it remains useful. It permits precise control over the differences between each ensemble, and allows a given ensemble to be placed in multiple acoustic environments. This in turn permits examination of the effect of environment suggested in the previous chapter as an important factor in the perception of some words. The synthesis methodology of adding and removing stops is described in section 6.1.1 above.

Although the synthesis method was different from that used in chapter three, as the equipment used was identical, the issues raised by listeners in the initial experiment can validly be considered here. Subsequent experiments used stereo samples to aid a perception of reality, but the initial experiment used mono samples due to the extreme left-right spread of the audio output from the experimental organ. This problem was solved by the use of an intermediate mixing desk that placed the left and right channels at 33% of their respective maximum pan distance.

The necessity of realistic synthesis meant that the reverberation unit built into the organ used in initial synthesis was no longer adequate. A Yamaha SREV 1 convolution based reverberation unit was placed between the mixing desk output and the computer audio input. The same experimental procedure used in previous chapters was followed, with four samples being compared in pairs and rated on scales for each adjective. The reasoning behind this methodology has been described in previous chapters, and there was no reason to change it for this final experiment.

The four ensembles had a number of different theories to test:

- Whether "thin" was related to the presence of higher harmonics or a lack of lower frequencies.
- Whether "flutey" continued its high correlation with a smooth ensemble and lower spectral centroid.
- Whether "warm" correlated with a lower spectral centroid, and whether reverberation had an effect.
- Whether "bright" continued to correlate with an increased spectral centroid.
- Whether "clear" correlated with any measurable feature.
- Whether all those words continued their apparent common understanding.

Given the limited choice of stops on the one keyboard still functioning at this stage and the necessity to keep this within the previously used context of the principal chorus, the synthesis of ensembles was a fairly simple affair. The following combinations of stops were chosen:

- Open Diapason 8', Principal 4', Fifteenth 2'
- Stopped Diapason 8', Principal 4', Fifteenth 2'
- Open Diapason 8', Principal 4', Fifteenth 2', Mixture IV

The first and third ensembles should have a difference in spectral centroid, with the Mixture ensemble being perceived as "brighter". The lower frequencies remain the same for both ensembles, but in the case of the second ensemble the lower frequencies are reduced, offering the chance of comparing both ways of increasing the spectral centroid. To examine the significance of reverberation, a fourth sample was recorded of the first ensemble in a more reverberant setting. The three initial samples used the preset "Warm Wooden Church no. 1" on the SREV unit, but the fourth sample used the preset "St John the Divine no. 3". The latter is more reverberant than the former by a large factor. There is no doubt that reverberation is more complex than a simple dimension such as RT_{60}, but the simple use here should indicate whether it is a factor in timbral perception worthy of future investigation.

Samples were presented to the listeners in files of less than 1MB, at uncompressed 16-bit stereo, 22.05kHz resolution, as discussed in section 3.1.3. The musical extract used was a small hymnodic stanza composed for the occasion by the author. Later analysis of the final chord by a custom Matlab script (presented in appendix I.2, and using an 8196 point Hamming windowed Fourier Analysis) permitted direct comparison of the four samples on three scales: centroid, slope, and smoothness. The formula for spectral centroid was the same as formula 5.1, although not normalised for frequency as there is no single fundamental in this instance. Spectral slope here is again the difference in average amplitude between the spectral energy in the two halves of the frequency spectrum. Spectral smoothness is a measure of average difference between one frequency bin (of which there are many more here than in the harmonic analysis) and the next. It was not possible to easily and accurately extract the other scales developed in chapter five. This limited analysis data is presented in table 6.1.

In all tables in this chapter, the organ specifications are abbreviated as follows:

- 842 is the basic principal chorus.
- S842 is the ensemble where a stopped diapason was substituted for the open diapason.
- 842M is the principal chorus to which a mixture has been added.
- 842 (rev) is the basic principal chorus placed in the more reverberant acoustic.

Sample analysed	842	S842	842M	842 (rev)
Spectral centroid (Hz)	1609	1665	2301	1459
Spectral slope (dB)	12.895	12.436	17.039	14.515
Spectral smoothness (dB)	2.047	2.143	2.752	2.116

Table 6.1 – Spectral centroid, slope and smoothness data for the four synthesised ensembles

The data in table 6.1 suggests that the alteration of the 8' stop used leads to a more subtle increase in spectral centroid. It is interesting that the alteration of environment results in a distinct reduction in spectral centroid, something that would not be apparent using a conventional reverberation unit that merely added reverberation to a dry signal. The SREV unit virtually places the dry sound at a location and by convolution calculates its effect at another location, those locations originally being those of a pseudo-random noise source and a microphone to pick up that noise (and its reverberation) respectively.

The limitations of the spectral analyses presented in figures 6.1 to 6.4 below have been described in section 6.1.1. Full amplitude data for these ensembles can be found in appendix K, but the figures concentrate on the first twenty harmonics to permit better comprehension of the data. Each graph shows all three ensemble's amplitude values for one of the four central voicing points used. The lowest voicing point had too much data to easily represent graphically and the highest is not relevant as it was not used in any of the audio examples. The amplitude data present in these graphs is prior to the addition of reverberation, so only three ensembles are shown.

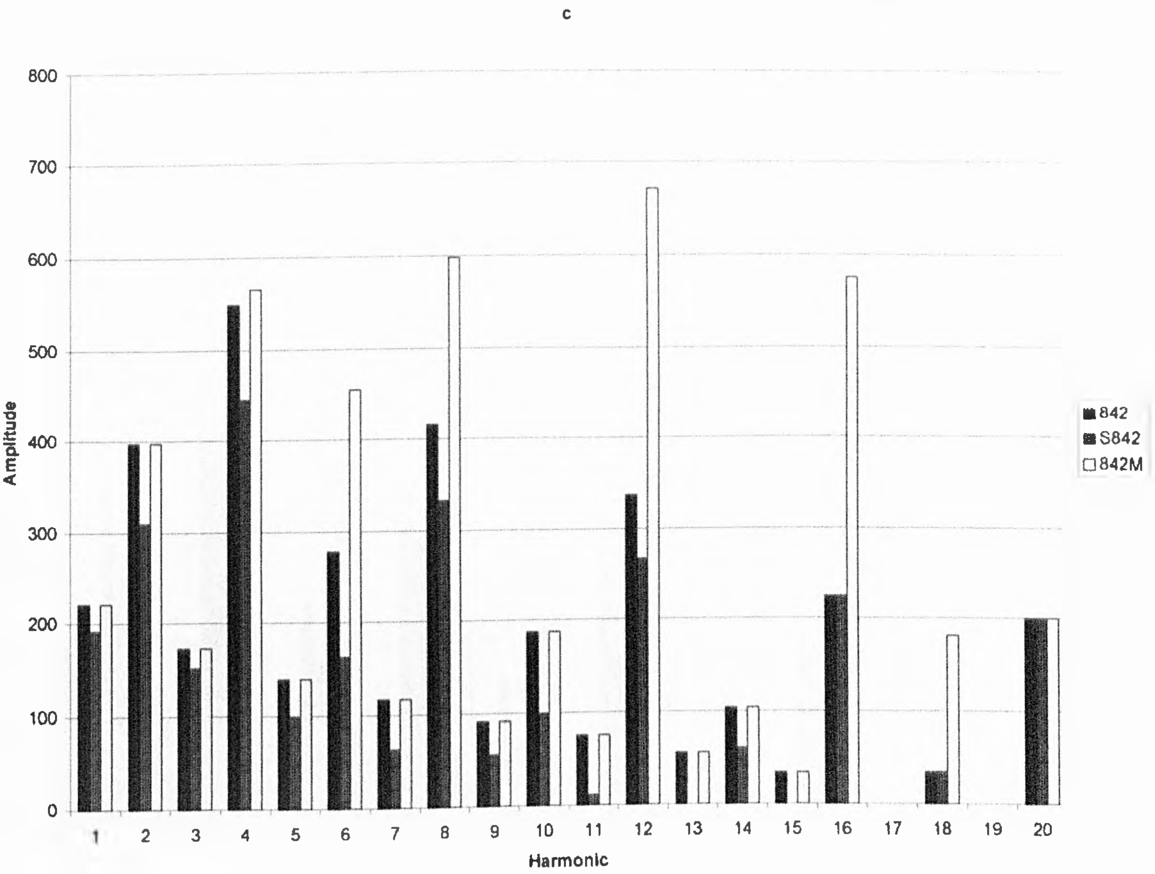


Figure 6.1 – Comparative amplitudes of the three synthesised ensembles for note 'c'

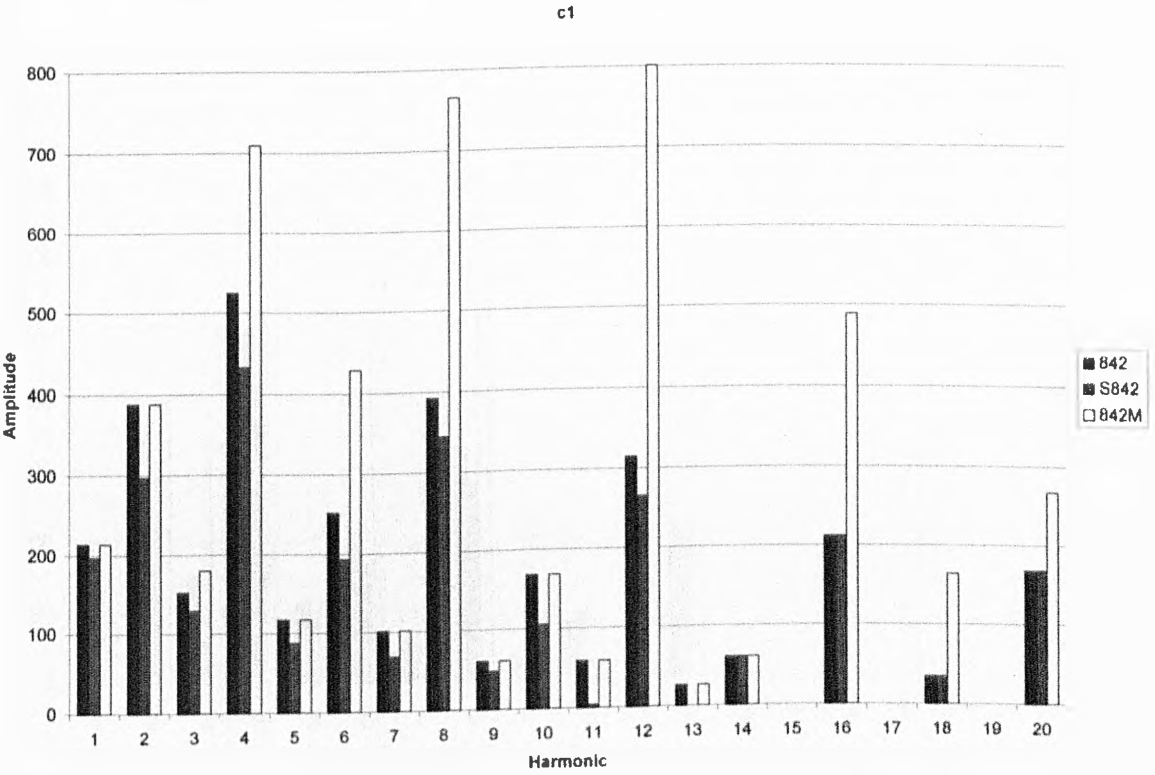


Figure 6.2 – Comparative amplitudes of the three synthesised ensembles for note 'c1'

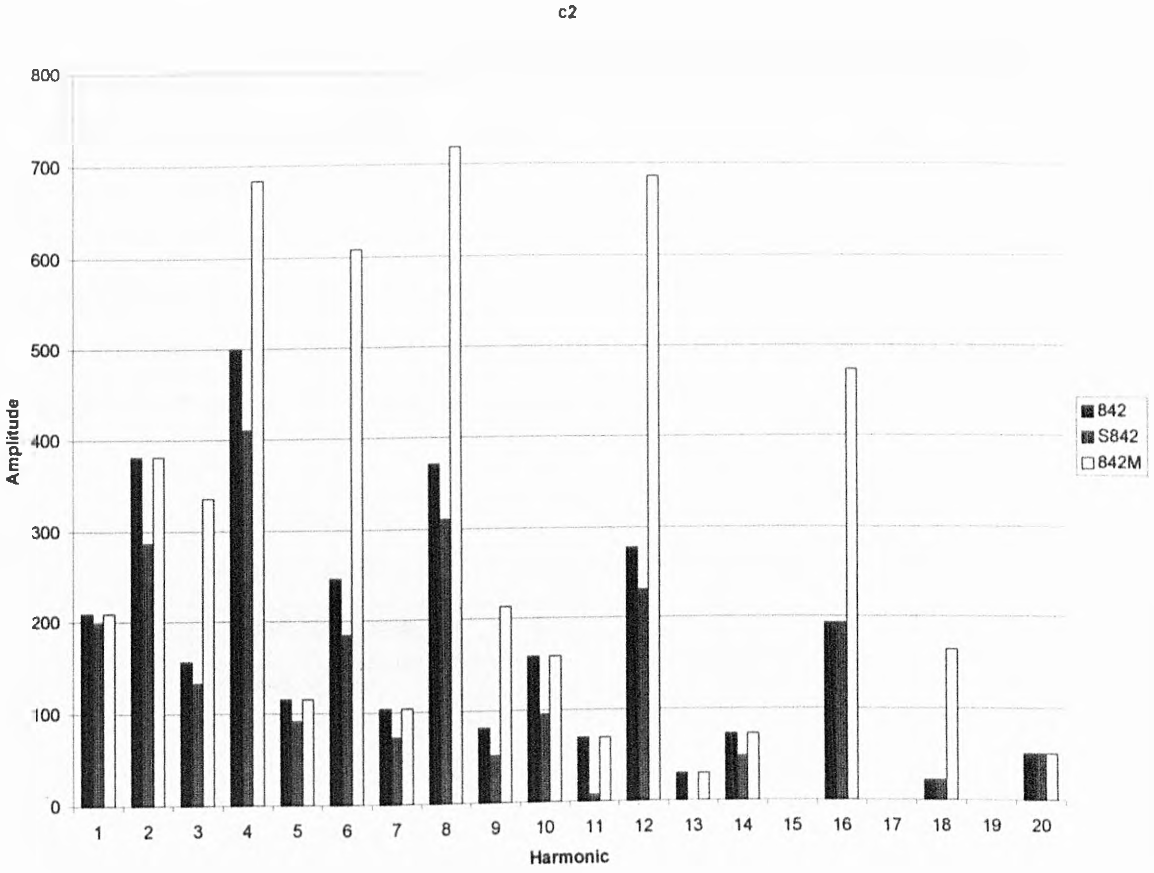


Figure 6.3 – Comparative amplitudes of the three synthesised ensembles for note 'c2'

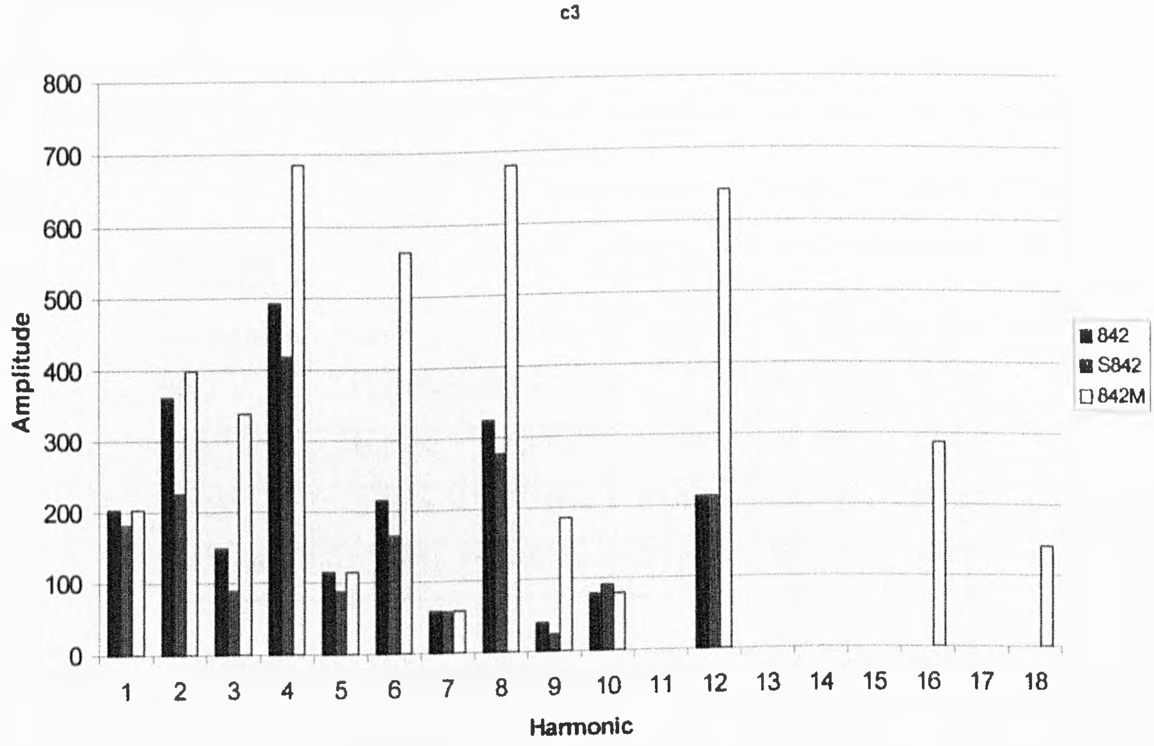


Figure 6.4 – Comparative amplitudes of the three synthesised ensembles for note 'c3'

A total of sixteen subjects took part in this final experiment, most of who had taken part in one or more of the earlier tests. While this number was smaller than for any of the other tests (apart from the initial experiment), it was still sufficiently large to be significant by the criteria laid out in Levetin (in Cook, 1999, p310). Age and geographical data for the subjects is given in table 6.2. One UK subject declined to give age data, so had his results excluded from any age-related analyses. Only one woman took part in this test, again making sex-related analyses impossible.

Geographic subject location	Age in years of subjects		
	Below 50	50 or above	Total
United Kingdom	4	3	8
United States	3	4	7
Canada	0	1	1
All subjects	7	8	16

Table 6.2 – Geographic and age data for listeners participating in final experiment

6.2 Presentation of results

The overall results for all sixteen subjects are presented in table 6.3 below. Detailed test by test results can be found in appendix L.

Organ ensemble	Adjective				
	thin	flutey	bright	warm	clear
842	-11.67%	12.92%	-17.92%	14.58%	-2.08%
S842	29.17%	2.50%	-7.50%	-8.33%	-0.83%
842M	-5.00%	-37.08%	57.08%	-32.92%	24.58%
842 (rev)	-12.50%	21.67%	-31.67%	26.67%	-21.67%

Table 6.3 – Overall results of final experiment for all sixteen subjects

There are several interesting results in table 6.3, but to give a firm basis for detailed analysis, it is necessary to examine the results for consistency of understanding. Tables

6.4 and 6.5 below present the results for young and old subjects respectively. As table 6.2 shows, young and old were roughly equally distributed among both main geographic groups, meaning that consideration of both differences is valid and not necessarily highly correlated.

Organ	Adjective				
ensemble	thin	flutey	bright	warm	clear
842	-17.14%	18.10%	-27.62%	14.29%	-4.76%
S842	27.62%	-0.95%	2.86%	-8.57%	7.62%
842M	11.43%	-40.95%	64.76%	-38.10%	15.24%
842 (rev)	-21.90%	23.81%	-40.00%	32.38%	-18.10%

Table 6.4 – Results of final experiment for the seven younger subjects

Organ	Adjective				
ensemble	thin	flutey	bright	warm	clear
842	-5.00%	6.67%	-8.33%	12.50%	-4.17%
S842	25.83%	5.00%	-13.33%	-5.00%	-7.50%
842M	-20.00%	-31.67%	49.17%	-25.83%	33.33%
842 (rev)	-0.83%	20.00%	-27.50%	18.33%	-21.67%

Table 6.5 – Results of final experiment for the eight older subjects

Here there are some interesting differences. While most answers appear to be in broad agreement, particularly noting the small sample sizes, there are some specific and significant disagreements. "Thin" has disagreement between both age groups, apart from strong agreement on the "S842" ensemble being the most "thin" by a large margin. "Flutey", "Bright" and "Warm" have the samples placed in the same order by both subject groups, with just the amounts differing. There is agreement as to which sample is most and least "clear", but not on the intermediate sample order. Pearson correlation coefficients provide a means of quantifying these disagreements while allowing for different interpretations of the scale ranges, and these are presented in table 6.6.

Adjective	thin	flutey	bright	warm	clear
Correlation coeff.	0.436	**0.971**	**0.945**	**0.992**	0.838

Table 6.6 – Pearson correlation coefficients for subjects divided by age

Significant correlation at the 5% level has been marked in bold. These results in table 6.6 confirm the analysis of the previous paragraph, although just because two subject groups do not correlate overall does not mean that the results for that adjective are insignificant. In the case of "thin", for example, subjects appear to agree on what is "thin" but not what is "not thin".

Examination of the second division of subjects, by geographic location, may shed further light on the level of agreement. Tables 6.7 and 6.8 present the results for subjects from the United Kingdom and United States of America respectively.

Organ ensemble	Adjective				
	thin	flutey	bright	warm	clear
842	-14.17%	10.83%	-23.33%	23.33%	0.00%
S842	35.00%	12.50%	-19.17%	-8.33%	-2.50%
842M	5.83%	-41.67%	70.83%	-41.67%	43.33%
842 (rev)	-26.67%	18.33%	-28.33%	26.67%	-40.83%

Table 6.7 – Results of final experiment for the eight UK subjects

Organ ensemble	Adjective				
	thin	flutey	bright	warm	clear
842	-9.52%	18.10%	-16.19%	5.71%	-3.81%
S842	25.71%	-11.43%	6.67%	-6.67%	0.95%
842M	-17.14%	-34.29%	44.76%	-25.71%	6.67%
842 (rev)	0.95%	27.62%	-35.24%	26.67%	-3.81%

Table 6.8 – Results of final experiment for the seven US subjects

Here again there is disagreement over which is the least "thin" ensemble, but not over which is the most "thin". "Flutey" has disagreement on the internal placing, but not on the extreme examples. "Bright" has agreement on the order, but the amounts differ – here it is interesting that the UK subjects use more of the scale than the US subjects – the opposite of what occurred in section 5.6.1. The overall absolute average for UK listeners in this part of the test was 24.6%, a dramatic increase from 15.04% in the previous test. The US absolute average was 16.05%, a slight drop from their previous 18.3%. "Warm" returns to broad agreement between the two subject groups, and "clear" shows both agreement on the extreme cases, but a dramatic difference in the magnitude of that agreement, with the US subject group overall showing little movement from the centre. Certainly the US listeners had a higher standard deviation for that word than the UK listeners. Overall average standard deviations are presented in table 6.9.

Subject group	Adjective					Overall average
	thin	flutey	bright	warm	clear	
All subjects	1.72	1.65	1.66	1.63	1.88	1.71
UK	1.52	1.66	1.33	1.81	1.51	1.56
US	1.78	1.52	1.80	1.44	1.88	1.68
Older	1.67	1.87	1.80	1.66	1.91	1.78
Younger	1.59	1.36	1.29	1.47	1.83	1.51

Table 6.9 – Average standard deviations for each subject group by word

The difference between the US and UK is not as great as that between the older and younger subject groups. The average standard deviation is slightly higher for those words that have less correlation between subject groups, but these standard deviations are less than for previous experiments in this thesis on average. This suggests that the way in which this research has focussed in on both words and spectral features has been successful.

Adjective	thin	flutey	bright	warm	clear
Correlation coeff.	0.603	0.838	**0.904**	**0.935**	0.860

Table 6.10 – Pearson correlation coefficients for subjects divided by location

Table 6.10 presents the Pearson correlation coefficients for the US compared with the UK. This demonstrates a slightly greater geographic difference on all words apart from "thin", where the age-related difference is greater. This is surprising, as in table 5.15 in chapter five, both sub-groups had more positive correlation for "thin", with the age-related difference being the lesser of the two.

Table 6.11 presents the average answer for each subject group, to verify that results have not been subject to skew.

Subject group	All	UK	US	Older	Younger
Average answer	6.04	5.99	6.08	6.18	5.91

Table 6.11 – Average answers for all subject groups

None of those results present any concerns of skew, given the small size of the sub-groups. It is also interesting to examine the correlation between individual words to see if the introduction of two new terms has resulted in any trends towards synonymy. Table 6.12 presents the average Pearson correlation coefficients between each adjective for all sixteen subjects. Correlation significant at the 5% level is marked in bold.

	Thin	Flutey	Bright	Warm	Clear
Thin	(1)	0.155	-0.084	-0.312	-0.146
Flutey	0.155	(1)	**-0.440**	0.301	-0.127
Bright	-0.084	**-0.440**	(1)	-0.357	0.349
Warm	-0.312	0.301	-0.357	(1)	-0.039
Clear	-0.146	-0.127	0.349	-0.039	(1)

Table 6.12 – Average Pearson correlation coefficients

The only correlation to reach significance is that of "flutey" and "bright" – it appears that the two are in opposition to a significant degree. The fact that no other pairs appear suggests that simple theories based entirely, for example, on the spectral centroid are inappropriate and more complex factors are at work. The tendency of several pairs towards correlation however suggests that spectral correlates such as the centroid could be responsible for a significant part of the auditory cues for use of a particular timbral adjective.

Unfortunately it is not possible to do any great analysis of the spectral features used in chapter 5 as very few of them could be extracted without the single note samples it proved impossible to obtain. This precludes the possibility of advanced analysis such as PCA even if it was desirable. Table 6.13 compares the three spectral features (centroid, slope and smoothness) presented in table 6.1 that had been obtained from the samples by the Matlab script in appendix I.2. Correlation significant at the 5% level is marked in bold.

Subject subgroup	Spectral feature	Adjective				
		thin	flutey	bright	warm	clear
All	centroid	0.002	**-0.994**	**0.999**	**-0.928**	**0.955**
	slope	-0.442	-0.740	0.769	-0.545	0.568
	smoothness	-0.082	**-0.958**	**0.966**	-0.857	0.846
US	centroid	-0.507	**-0.904**	**0.954**	**-0.903**	**0.943**
	slope	-0.698	-0.522	0.584	-0.440	0.677
	smoothness	-0.520	-0.841	0.880	-0.783	**0.926**
UK	centroid	0.333	**-0.990**	**0.987**	**-0.934**	**0.937**
	slope	-0.209	-0.852	0.868	-0.631	0.534
	smoothness	0.213	**-0.979**	**0.991**	**-0.903**	0.813
Younger	centroid	0.484	**-0.978**	**0.976**	**-0.934**	0.839
	slope	-0.005	-0.715	0.693	-0.535	0.333
	smoothness	0.397	**-0.950**	**0.939**	-0.851	0.712
Older	centroid	-0.576	**-0.996**	**0.993**	**-0.943**	**0.988**
	slope	-0.807	-0.732	0.785	-0.605	0.727
	smoothness	-0.623	**-0.940**	**0.949**	-0.893	**0.923**

Table 6.13 – Pearson correlation coefficients between spectral features and adjectives

It is interesting to note that the Pearson correlation coefficient between the values for spectral centroid and slope is 0.778, indicating some positive correlation but below the level of statistical significance for four samples at the 5% level. The coefficient between smoothness and spectral slope is 0.898, almost at significance, and that between spectral centroid and smoothness is 0.965, well above significance. It is therefore not surprising that those two terms produce such similar results.

The extremely high overall correlation between most of the adjectives and the spectral centroid is interesting, as previous analysis and discussion has suggested that the four adjectives are not strongly correlated. Clearly the centroid is an important factor in the use of certain words, but its significant correlation with many words in table 6.13 does not mean it is the sole factor in determining their use. Most significant is "bright", which together with "flutey" would seem to have the closest relationship to the spectral centroid. Spectral slope lacks significant correlation with any adjective for any subject group. Spectral smoothness, due to its high correlation with the spectral centroid, follows that factor in all cases though to a slightly lower level of significance. The correlation between "clear" and the spectral centroid dips below statistical significance for the younger subjects group, suggesting that while important, other factors are at work for this group. Younger subjects did not have the highest standard deviation for "clear", so this dip is not due to disagreement between the listeners in that group. The high levels of correlation between the spectral centroid and most adjectives make it all the more regrettable that it was not possible to derive more spectral features from the synthesised samples.

Subjects were told the nature of the test samples immediately afterwards, and their comments suggested that a greater illusion of use of real samples was maintained, and the different reverberant environment meant that the two identical registrations were not thought alike.

6.3 Analysis and discussion

Although it has not been possible to carry out all the synthesis and analysis originally desired in this section, this should not detract from the significant results gained using a simpler form of synthesis, which also brings those results nearer to the real-world possibilities of a typical pipe organ. The theories and results relating to each of the remaining adjectives will be considered individually in this section.

6.3.1 "Thin"

While "thin" was related to a higher spectral centroid, it was unclear whether this was due to the presence of higher harmonics, as suggested in the previous chapter, or the lack of lower harmonics, as suggested by previous research. The results presented in table 6.3 suggest very clearly that the reduction in lower harmonics is responsible for perceived "thinness". When the mixture was added, despite a much greater increase in the spectral centroid, the ensemble was perceived as less thin.

There was not universal agreement on this aspect of "thinness" by listeners. Although all agreed on the lack of lower harmonics being responsible for "thinness", there was some disagreement between the different subject groups as to whether the presence of higher harmonics contributed to "thinness". All subjects were however agreed that the lack of lower harmonics was far more important. Overall, "thin" did not significantly correlate with the spectral centroid, suggesting that the observations in chapter five noting its tendency towards correlation with the centroid was just a factor of the particular organ ensembles chosen for use in that part of the research. Similarly, it lacked the strong correlation in the previous chapter with an unsmooth ensemble, suggesting that care must be taken when basing theories on only one piece of evidence.

6.3.2 "Flutey"

"Flutey" is probably not a word that springs to mind as one of the most useful descriptive terms for a pipe organ ensemble, and indeed has no previous research, but it has retained common understanding throughout the course of this research. "Flutey" retains its negative correlation with spectral centroid and spectral smoothness, although the two have exchanged places in the order of most significant compared to the results of the previous chapter. While both centroid and smoothness ratings overall are correlated, previous chapters have demonstrated that highly correlated spectral features can sometimes shed light on the precise triggers for certain words. In this case, the exchange in positions suggests that more work is necessary using more ensembles to pinpoint the precise spectral trigger for "flutey", but more significant in the context of this thesis is the fact that "flutey" has retained its common understanding throughout.

6.3.3 "Warm"

"Warm" continues to correlate with a lower spectral centroid, but to a lesser extent than "flutey". This relationship with the centroid is consistently significant for all subject subgroups. "Warm" also reaches significance with some subgroups for spectral smoothness, but this is on average below a statistically significant figure. The suggestion was made in chapter five that one possibility for the difference between "flutey" and "warm" was the amount of reverberation present. The effect of reverberation will be considered in more detail in section 6.3.6 below, but for the US subjects the addition of reverberation added to their perception of the sound as "warm" more than their perception of it as "flutey". Subjects from the UK matched those from the US in their "flutey" ratings, but did not feel that the added reverberation contributed to an increased perception of "warm".

"Warm" was related in the previous research described in section 5.1 to a decrease in spectral centroid and an increase in the strength of the first few harmonics. This latter quality is impossible to measure from the complex chords analysed in section 6.2, but in the test directly comparing the 842 and S842 samples, of which the latter had reduced lower harmonics, the former was considered warmer overall by an average of 15%. The theory of decreased spectral slope also mentioned in section 5.1 is supported inasmuch as the correlation between slope and perceived warmth is consistently negative, but well below statistical significance.

6.3.4 "Bright"

In contrast to "warm", "bright" maintains a very clear correlation with the spectral centroid throughout all subject subgroups and indeed all tests in this thesis. In this final test the overall correlation is very high, and "bright" also seems to a significant degree to be an opposite of "flutey" in this context. It could be interesting to see if "flutey" was essentially being used as a synonym for "dull", one of the less common words gathered but ultimately rejected for experimentation in chapter four. While this result matches the previous theories described in section 5.1, it has not been possible to more precisely isolate which theory applies in the case of the pipe organ due to the problem in discerning between spectral width and centroid in this context.

6.3.5 "Clear"

"Clear" in the previous chapters appeared not to follow previous theories, although this was one of the two words not to be subject to the more rigorous specialised listener analysis. In this chapter's experiment a statistically significant correlation emerges with a higher spectral centroid, which broadly matches some of the previous theories described in section 5.1. This correlation is maintained by all subgroups apart from the younger listeners, for whom it is just below significance.

Ethington and Punch (1994) suggest that clarity relates to a decreased number of harmonics and expanded harmonic density. This does not appear to be supported by these results. The theories of Solomon (1959), relating clarity positively to high frequency energy, neutrally with mid frequencies, and negatively to low frequencies appears to be supported by the results of this work. Some insight into this theory and that of Jeans (1961) can be gained by looking at the results of the test directly combining the 842 and S842 examples, the latter of which had a distinctly lower amplitude second harmonic and less amplitude for the lower harmonics in general. Here there was little difference in the perceived clarity, suggesting that at least in the context of the pipe organ clarity is not related to the strength of the lower harmonics.

The addition of a mixture improves the perception of "clear", but there was a geographical difference on the amount to which that perception is improved, with the UK subjects giving it stronger support than those from the US. There was also geographical difference on the effect of reverberation. When the single test comparing the 842 samples with and without reverberation was taken, UK subjects on average thought that reverberation decreased clarity, but the US subjects did not.

6.3.6 The effect of reverberation

As has been mentioned previously in sections 6.3.5 and 6.3.3, the addition of reverberation is significant in the use of certain timbral adjectives, but before the effect of reverberation can be looked at, the nature of reverberation must be considered in more detail. Reverberation, the way in which sound is reflected and shaped by the acoustic environment in which the sound is placed, is not a simple linear quantity as might be implied by such measures as RT_{60}, but can also have a spectral filtering element. Thus in the example of increased reverberation used earlier in this chapter, there is a marked reduction in the spectral centroid and an increase in the spectral slope. It is entirely possible that another reverberant environment would provide an equivalent increase in RT_{60} without the same spectral effect, due to the presence of physically and acoustically different sound-reflecting and absorbing materials. However, the aim of this particular experiment is to examine whether a difference in reverberation has an effect rather than to precisely quantify that effect.

Within those constraints, the effect of the alteration in reverberant environment on the timbral adjectives studied in this section can be examined. "Thin" was the subject of some disagreement, resulting in an overall neutral result that concealed the differences of opinion. As might be expected from the reduction in spectral centroid of the reverberant example, this was perceived as less "bright" than the less reverberant example. However it was interesting to note that the US subjects felt this more strongly than those from the UK. As described in section 6.3.3, US subjects also felt that the addition of reverberation added to a perception of "warm", whereas UK subjects felt it added to a decreased perception of "clear". Both these adjectives have a relationship to the spectral centroid that does not make these perceptions unrelated, but the geographical difference of emphasis is interesting. Both subject groups agreed on an increase in the perception of "flutey" when reverberation was added.

6.3.7 Common understanding of timbral adjectives

The common understanding indicated by the results of chapter five continues to be demonstrated by the results of this chapter. The subdivision of subjects into geographic and age-related groups has permitted examination of how this understanding varies, and on average there is less difference between the two age groups than between the two geographic groups. "Bright", "warm" and "flutey" all exhibit common understanding that is reasonably universal. "Thin" has a degree of common understanding, but there is disagreement on what is more "not thin". "Clear" has some common understanding, but that understanding appears different for differing nationalities, with the UK subject group being far more definite in their answers.

6.4 Summary of theories developed

Within the context of English speakers describing the principal ensemble of a pipe organ, listeners describing a sound as "thin" will be referring to a comparative lack in strength of the lower harmonics. A "flutey" ensemble will have a lower spectral centroid, as will a "warm" ensemble. A "bright" ensemble will have a higher spectral centroid, as will a "clear" ensemble.

Such an ensemble placed in a more reverberant environment may be considered more "flutey" and "warm", the latter particularly the case with US listeners. UK listeners thought that the more reverberant environment made the ensemble less "clear".

Most importantly within the context of this thesis' hypothesis, common understanding of timbral adjectives has continued to be demonstrated. Listeners' perceptions of "thin" and "clear" have demonstrated how significant common understanding does not preclude some differences both among and between the subgroups of subjects.

7. Conclusion

The previous chapters in this thesis have described in detail the aims of this thesis and the methodology used to implement those aims, along with the results for each experimental stage. This chapter brings all that work together, with a brief overview of the research programme, and a summary and discussion of the theories developed. There is a brief summary of possible future work that could arise from different sections of this thesis, then the final section re-examines the hypothesis and lists the most significant novel aspects of this research.

7.1 Overview of research conducted

Chapter one expanded the concept described in the hypothesis to a three-stage process: to gather words use by people to describe the sound of the pipe organ, then look for common understanding of those words between listeners, and then to look for possible correlation with spectral features. The particular usefulness of the pipe organ as a tool for psychoacoustic experimentation was described, and timbral semantics as an area was introduced. The reasons behind the use of the principal ensemble and the importance of using a significant proportion of real samples were elaborated.

A programme of research was proposed: to look at existing work in the field of timbral semantics, to conduct some initial experiments, to gather and refine timbral adjectives, and then examine them as rating scales for audio examples, looking for common understanding and ultimately spectral correlation.

Chapter two examined all relevant recent research known to the author, and chapter three implemented a pilot experiment based in part upon the author's previous research. That experiment suggested real but variable common understanding of timbral adjectives, and provided useful guidelines for the other experiments in later chapters.

Chapter four sought to gather and select timbral adjectives by means of a free-choice response to several audio examples. The opportunity was also taken to study some related factors such as the effect of displayed images on listener responses.

As a result of various selection procedures, seven adjectives were selected for further study based mainly on frequency of occurrence. Chapter five refined these further to five adjectives, all of which demonstrated a significant degree of common understanding. Chapter six sought to both verify that common understanding for those words and to quantify it in terms of spectral features that showed some correlation with the occurrence of those timbral adjectives.

Chapter five used recordings of real organs, whereas chapter six used synthesised ensembles designed to be as realistic as possible, based on earlier listener criticism of synthesised examples as obviously artificial. This synthesis was limited in its scope by problems with the experimental equipment, but this did not prevent the verification of common understanding and the establishment of some correlation between timbral adjectives and spectral features.

At the time of thesis submission, two papers had been published with results directly arising from this piece of work. Disley and Howard (2003) summarised the main research, and Disley and Howard (2004) examined the spectral correlation in more detail. An additional article in an organ journal (Disley, 2003) described the study in chapter four of listeners' reactions to different combinations of picture and sound samples.

7.2 Summary and discussion of theories developed

The hypothesis of this thesis had two main goals: to establish if there was pre-existent common understanding of timbral adjectives, and if so, to examine the nature of that understanding, both in the context of the pipe organ. The initial aim has been proved, with common understanding demonstrated by the results of several chapters, but differing levels of understanding were demonstrated for different adjectives, and differences were evident among the listeners when they were divided into groups by age and location. The second aim was developed as the thesis progressed.

In chapter three, the difference between a agreeing on a common definition for a timbral adjective and agreeing on an interpretation of that in the psychoacoustic domain was readily illustrated. Listeners demonstrated no consistency when describing the blend of an ensemble, but appeared to agree on the definition of "blend" given to them. This demonstrates the complex nature of this field, and the need not to rely on the apparently obvious even when many of the later results do seem to match what might be expected by "common sense".

Chapter three also introduced the possibility of listeners assuming that the words on certain scales were associated with a good ensemble. A significant correlation was observed between older subjects' ratings for "blend" and preference. Subsequent tests including a preference scale in chapter four did not appear to support this hypothesis, but the possibility remains that certain words might be associated with a good organ in subjects' minds. Thus a subject might think "I prefer that organ, therefore it must be better blended, as blend is a sign of a good organ", rather than critically examining the ensemble in question for perceived blend. It is difficult to isolate such terms, as all good organs might be good because their choruses are well blended. Ultimately, such relationships can only be commented upon, as it is unwise to recommend avoidance of any words that could correlate with preference on the basis of one such possibility.

Chapter five went on to develop theories of spectral correlation for five words commonly used by subjects in chapter four to describe the sound of the pipe organ. Two words were rejected as they fell below the threshold of demonstrable common understanding. Correlation was used extensively in chapters five and six as a means of demonstrating common understanding, but while a useful statistical technique, it cannot be relied upon as a sole means of proving or rejecting a certain theory. As the theory for the word "thin" demonstrated, lack of correlation can hide results that remain significant outside of the concept of statistical significance. However, the two rejected words, "balanced" and "full", did not appear to have any common understanding even outside of the constraints of statistical analyses. It is entirely possible that those words have common understanding in the wider context of the pipe organ, but in the specific context of the principal ensemble, despite their common usage, these words are not commonly understood. This has interesting consequences for people who use timbral adjectives, as it suggests that just because a word is in common use, different people may still have different understandings, and thus frequency of occurrence cannot be sole grounds for use of a particular term.

The theories developed in chapter five were presented in section 5.8 and refined in section 6.4. Although a wide variety of spectral analyses were developed suitable for use in the context of ensembles (rather than single pipe samples), it is disappointing that most theories essentially related to the spectral centroid. This is partially due to the fact that many of the analyses developed in chapter five were impossible to implement in chapter six due to the failure of experimental equipment. However it is also possible that many of the analyses developed might have less correlation with each other outside of the limited scope of this thesis. It would also be useful to develop further analyses suitable for multiple source ensembles, including those that could permit time-related analysis that did not ignore the collective ensemble transient.

The specific theories relating common timbral adjectives to spectral can be briefly summarised. A "thin" sound relates to a lack of strong lower harmonics, "flutey" and "warm" sounds have lower spectral centroids, and "bright" and "clear" sounds have higher spectral centroids. These theories can also be expressed in terms of the effect of stop manipulation by an organist. If a mixture is added to an ensemble, this will add to listeners' perceptions of it as "bright" and "clear", and there will be a decrease in perception of it as "warm" and "flutey". If a Stopped Diapason replaces an Open Diapason in an ensemble, the result will be perceived as more "thin" and slightly less "warm". Organists tend not to think in terms of what words or scales their stop manipulations will achieve, but instead follow either less easily defined processes of instinctive decision making or pre-determined patterns from training or musical direction. The ability to respond accurately to third-party requests to make an ensemble "less bright", for example, is one that may not be welcomed by all organists, but could be useful in the context of ensemble playing where the organ must blend with other instruments.

As many of the potential spectral correlates have not been examined beyond their initial use in chapter five, it is difficult to comment on the usefulness of measures such as the variation of certain qualities over the keyboard range. Although some of these were not correlated with other spectral features, this alone does not imply that they are useful. The correlation of all features needs examining against a greater variety of sample sources before any can be rejected as consistently correlating with another feature. It is difficult in the case of some clearly different but correlated spectral features to determine which is responsible for a particular perceptual quality. It may be, for example, that within the context of pipe organs it is impossible to significantly raise the spectral centroid without reducing the smoothness of the overall ensemble. However, it may be that many of these potential ensemble analysis scales are applicable outside the context of the pipe organ, and here some correlation might disappear.

The effect of different reverberant environments on an identical source was explored briefly in chapter six. This was not an attempt to cover that topic in any detail, but to begin to examine how different reverberant environments might affect both the sound itself and the words applied to it. The geographical differences evident there were interesting, and continued the trend evident from the initial experiment in chapter three, confirming the author's suspicions from section 1.5.2 that even listeners with a common language (English in this case) could have different understandings.

It is easy to compare the subtle (or not so subtle) differences in understanding between groups such as US and UK English speakers with the known differences in language that result in some words having differences in emphasis or outright meaning. This could be a strong contributing factor to the different answers between subgroups for some adjectives. However, differences in test procedure might also be responsible for the differences in some words, despite all participants using the same interface. The use of comparative tests and the checking of results for bias were meant to ensure that the inevitable differences in testing environment for remote subjects did not affect the results. It is possible that on average the US listeners used audio equipment less able to reproduce the subtle differences between samples, but informal discussions with the subjects suggested that this was not the case. It is also possible that as most of the UK subjects knew the author personally, they had a different understanding of what was required of them. However, this would suggest that all of the results should have shown greater geographical variation instead of just some results, and as all listeners were informed of the test aims as the test progressed, this would imply that US subjects consistently ignored those instructions. This does not appear to have been the case. The conclusion on this issue remains, therefore, that linguistic differences between subjects from different geographical locations with a nominally common tongue can play a significant role in the understanding of some timbral adjectives.

This has some interesting consequences for future studies, as many previous studies the author has referenced (e.g. Sandell (1991) and Rioux (2001)) did not include any geographic data on their subjects. In the latter case this was particularly unfortunate, as Rioux's subjects were attendees at a conference from many different countries, many of whom would have spoken English as a second or third language. Other studies, such as Nykänen and Johansson (2003) or Moravec and Štěpánek (2003), have concentrated on a single linguistic group, either deliberately, or accidentally through their conducting of the tests in their native tongue. Future such tests must be careful to indicate the native adjectives as of primary importance, and note that English translations used in papers may have slightly different understanding among listeners.

Perhaps the most important result from this thesis within the context of the hypothesis is the continued demonstration of listeners' common understanding of certain timbral adjectives. Listeners' perceptions of "thin" and "clear" have demonstrated how undoubtedly important common understanding of some aspects of a term does not equate to universal agreement on all aspects, both among and between the participant subgroups.

Psychoacousticians have in the past attempted to define certain core sets of adjectives, often pairs in opposition, that adequately describe the majority of timbre space. Examples include Pratt and Doak (1976) and von Bismarck (1974). Creasey's opposition to such reductionist trends has already been covered in chapter two, but it is interesting to note that the only term to appear in either of those examples and the words selected for examination, "full", was rejected as lacking common understanding in this context. It was not presented in the opposition to "empty" as in von Bismarck, but the lack of any other such terms here suggests that listeners have different priorities when describing sounds. Although the terms used by listeners may be closely correlated within a context, they have the ability to describe subtleties of timbre lacking in more general descriptions. Listeners may also use specific subsets of timbral adjectives to describe the sounds of a certain instrument (the pipe organ in this case) and also its own different subsets of timbre (such as the principal chorus in this case). This ability to adjust the terms of reference suggests that while more general scales may be appropriate where the timbre space is large, when a specific timbre space is being used, those scales should be altered to ones appropriate for both timbre and listeners involved.

It is also important to note that where people make use of the words used on some of those scales, either defined generally or for a specific context, they do not necessarily have the same meaning or understanding by the listeners as their formal definitions. This could mean that for each series of experiments, it is necessary not only to establish an appropriate set of rating scales for the context, but also to examine the precise understanding of those rating scales by, for example, looking for spectral correlates as in this thesis. Ultimately it may be that such studies follow a pattern of diminishing returns, and given the very complex nature of timbral semantics such studies should only be carried out where there is an obvious use for the results.

One of the aims developed in this thesis was to remain as much as possible within the context of the pipe organ, meaning an avoidance of obviously artificial examples and the use of real samples wherever possible. Creasey (1998, p43) suggests that synthesised samples can be unsatisfactory as the complexity of traditional instruments is rarely matched by synthesised examples. However, the samples produced for use in chapter six were created on a very specialised synthesiser and processed via one of the most advanced reverberation units available. Listener comments included "I wondered how you had been able to record such obviously important instruments", presumably referring to the usual difficulty of getting access to the most significant pipe organs. The use of such specialised equipment has permitted the detailed examination of the spectral correlates in a way difficult to achieve with real pipe organs. Creasey's concerns are not shared by all studies, with some mentioned in chapter two suggesting that the use of synthesis may not be important from a timbral point of view. Certainly the careful use of synthesis, following the analysis and synthesis model of Gabrielsson (1985) appears to have worked successfully in this instance, with a gradual reduction in average standard deviation as the experiments progressed.

This could of course be the result of other factors, such as increased subject precision, caused by involuntary and unintentional training during the course of experimental procedure. However, this would assume a consistent subject group, and while some participants did take part in all experiments, others did not. At each stage of experimentation, some listeners were new and others had taken part in similar experiments by the author either as part of this thesis or previous work.

Previous studies on timbre have tended to rely on PCA or MDS to define a limited set of scales by which the samples used can be defined. The problems with both methods is that they remove the results from an immediate relationship with quantifiable reality. PCA could reduce, for example, the many potential spectral correlates to just a few, but those few would still be defined in terms of all the other correlates, gaining little in real terms within the context of this thesis. Each experiment lacks sufficient samples for a good MDS, but even if more samples had been used, MDS can make accurate general description of the constituent parts of timbre almost impossible in a concise manner (Creasey, 1998, p19). Any attempt to establish comprehensive axes of timbre space would inevitably be more complex than previous work given the multi-source nature of pipe organ ensembles even if sufficient samples were available.

There is a risk even in using the timbral adjectives gathered and refined in chapter four that subjects may assume that those scales are both uncorrelated and comprehensive in their description of timbral space (Creasey, p59). The specific instructions to subjects sought to avoid this possibility, but that risk remains in any similar study.

The five words that have demonstrated degrees of common understanding throughout this thesis ("bright", "clear", "flutey", "thin" and "warm") are not intended as any kind of complete description of this timbre space, but are interesting to consider in that light given the correlation between some of them. Of those five, "flutey" is perhaps the least expected, and it would be interesting to compare all of them with other adjectives such as those suggested for more general use to see if further correlation was evident.

In all the theories discussed and developed in this thesis, there is no suggestion that the spectral correlates suggested are the sole trigger for a particular adjective, but within the constraints of this study evidence suggests that those theories are worthy of further examination. Any study is limited by the samples it chooses, and the limitations imposed here by subject location and ability have necessitated a restricted sample set and hence restricted timbre space. This does not necessarily mean that the theories are only applicable to this timbre space, and the possibility of wider applicability should not be ignored.

7.3 Further work

This thesis provides a number of results that could be the starting point of future research, both in its main focus and in several subsidiary sections.

Section 4.5.7 considered the significance of displaying different pipe organ images while playing the same sample and concluded that there was some effect upon the words the subject then used. The particular combination of picture and sound presented was remembered by some subjects for significant periods of time, whether that combination was correct or not. Further research could attempt to quantify both these effects in a way that was impossible in the scope of this research.

Chapter four offers an interesting linguistic question on how adjective classification is affected by context, and chapter five presents more questions on the nature and effect of musicality on listener judgement. While musicality has been studied for some time (for details see the summary in Shuter-Dyson, 1999), its applied effect is less known. Its effect, even when the classification is self-conducted, is clearly important as the results from section 5.4 show. It would be interesting to compare those with a more precise experiment in which both subject musicality and their responses on semantic scales were examined.

The questions of what has caused the significant differences demonstrated between geographical and, to a lesser extent, age subgroups of listeners, have not been examined in detail by this thesis. Similarly, while common understanding has been demonstrated, the causes of that understanding have not been investigated. Some research has been done into how relationships between emotional states and music are established (summarised in Dowling, 1999), but the more complex nature of timbral adjectives' origins has not yet been explored. The results of this thesis would suggest both that such adjectives have been learnt, because of their common understanding, and that that learning process is localised, because of the geographical differences evident. Age-related differences could be explained by subjects' different hearing abilities or different early exposure to particular types of pipe organs, but as this thesis has not explored that area in detail it would be foolish to speculate further.

Several issues surrounding the choice of examples and recording methods offer potential for further study. All organs studied were nominally in equal temperament, but temperament could contribute to differences in listeners' perceptions of a pipe organ. Section 6.3.6 indicates that the building acoustic plays an important role in shaping the sound presented to the listener, but the perceptual role of both reverberation and sound dispersal in a room have not been studied further. Both could alter listeners' perception of the size of an ensemble as well as its timbral qualities.

While the reasoning behind the use of limited remote testing has been covered in some depth in this thesis, precise consideration of the effect of remote testing on subjects and their results would be very useful, particularly in enabling specialised experiments to use a dispersed subject set. Factors such as the quality of equipment used by such remote subjects need to be considered.

Finally, all analyses in chapters five and six concentrated solely on the steady-state portion of the sound. There is no doubt that starting transients are perceptually significant, and work identifying analyses appropriate for multiple-source ensembles would be of great use in unpacking their effect on timbral semantics.

7.4 Thesis significance and hypothesis testing

This thesis provides a significant and original contribution to understanding in the field of timbral semantics, being novel in both scope and methodology. In scope it has set new ground in considering listener responses to real, multiple source samples. Previous research on timbral semantics, in the rare cases where the whole field had not been written off as inherently subjective, has concentrated on single-source anechoic samples, which do not represent listeners' typical experiences of those sounds. This thesis has moved that research into sounds and ensembles that are found in the real-world, and the results suggest that despite the distinct increase in complexity, significant relationships between adjectives and measurable phenomena can still be identified.

In methodology, the use of an indirect method in chapter four to identify words actually used by listeners (as opposed to those known to them) was unprecedented in the field, as was the level of reality gained from the synthesis and reverberation system used in chapter six. The secondary consideration of usefulness (section 4.6) is not known to have been used elsewhere. Much other work, such as the significance of displayed images (section 4.5) and the significance of musicality (section 5.4.1), was novel in its application to this field. The timbre space used in this experiment was previously unexplored in this manner.

This thesis provides a number of contributions to knowledge. The methodology developed and used throughout the thesis can easily be applied to other experiments in a wide variety of related fields, providing a rigorous framework for psychoacoustic experiments. The use of Internet testing in this thesis could provide useful experimental alternatives where specialised subjects are required, provided that care is taken in its implementation.

Perhaps the most significant contribution to knowledge are the results refined in chapter six. These suggest that there are consistent relationships between certain timbral adjectives and acoustic phenomena, which indicates that timbral adjectives not entirely subjective in nature as suggested or assumed elsewhere in the literature. This in turn inspires a whole range of possible future research to further explore this relationship, both examining different timbral areas and looking at how the relationship is developed, as described in section 7.3. These results suggest that similar relationships are likely to be found in a large range of other fields. The hypothesis of this thesis is:

> Within the timbre space of the pipe organ, some timbral adjectives can be shown to have common understanding and a consistent correlation with acoustic phenomena

This hypothesis is supported by the results of the three questions asked in section 1.1. Five of the most common words used by people to describe the sound of the pipe organ, gathered and selected in chapter four, were shown to have a consistent understanding across multiple listeners in chapters five and six. Several auditory cues demonstrated consistent correlation with those words, and those theories are summarised in section 6.4. Thus within the timbre space of the pipe organ, some timbral adjectives have been shown to have common understanding and a consistent correlation with acoustic phenomena.

Appendix A: Numerical data from section 3.2

A1: Base specification

Harmonic	c	c1	c2	c3
1	221	214	209	203
2	199	190	197	185
3	173	152	156	150
4	154	137	138	124
5	139	111	109	105
6	125	99	101	73
7	99	83	86	39
8	119	100	116	64
9	74	37	49	10
10	64	41	40	
11	48	32	37	
12	40	21	9	
13	24	0	0	
14	15			
15	8			
16	40			

Table A.1 – Open Diapason harmonic amplitude data

Harmonic	c	c1	c2	c3
1	198	198	185	178
2	187	187	174	167
3	159	159	152	123
4	142	136	125	86
5	97	99	74	51
6	84	90	67	30
7	43	39	26	
8	100	94	81	
9	7	9	0	
10	31	31	22	

Table A.2 – Principal harmonic amplitude data

Harmonic	**c**	**c1**	**c2**	**c3**
1	188	183	181	179
2	167	175	151	146
3	130	127	115	104
4	116	108	87	71
5	85	82	35	33
6	61	61	13	0
7	51	47	0	
8	82	73	40	
9	34	17	0	
10	34	13	0	
11	12	11		
12	10	8		
13	14	9		
14	0	0		

Table A.3 – Twelfth harmonic amplitude data

Harmonic	**c**	**c1**	**c2**	**c3**
1	208	201	188	196
2	191	178	132	104
3	153	140	86	41
4	129	102	20	
5	109	79		
6	91	51		
7	68	25		
8	99	49		
9	39	0		
10	47	0		
11	34			
12	42			
13	26			
14	32			
15	19			
16	54			
17	0			
18	0			
19	0			
20	0			

Table A.4 – Fifteenth harmonic amplitude data

A2: Dull specification

Harmonic	c	c1	c2	c3
1	221	214	209	203
2	199	190	197	185
3	173	152	156	150
4	154	137	138	124
5	139	111	109	105
6	125	99	101	73
7	99	83	86	39
8	79	60	76	24
9	74	37	49	10
10	64	41	40	
11	48	32	37	
12	40	21	9	
13	24	0	0	
14	15			
15	8			
16	0			

Table A.5 – Open Diapason harmonic amplitude data

Harmonic	c	c1	c2	c3
1	198	198	185	178
2	187	187	174	167
3	159	159	152	123
4	142	136	125	86
5	97	99	74	51
6	84	90	67	30
7	43	39	26	
8	60	54	41	
9	7	9	0	
10	31	31	22	

Table A.6 – Principal harmonic amplitude data

Harmonic	c	c1	c2	c3
1	188	183	181	179
2	167	175	151	146
3	130	127	115	104
4	116	108	87	71
5	85	82	35	33
6	61	61	13	0
7	51	47	0	
8	42	33	0	
9	34	17	0	
10	34	13	0	
11	12	11		
12	10	8		
13	14	9		
14	0	0		

Table A.7 – Twelfth harmonic amplitude data

Harmonic	c	c1	c2	c3
1	208	201	188	196
2	191	178	132	104
3	153	140	86	41
4	129	102	20	
5	109	79		
6	91	51		
7	68	25		
8	59	9		
9	39	0		
10	47	0		
11	34			
12	42			
13	26			
14	32			
15	19			
16	14			
17	0			
18	0			
19	0			
20	0			

Table A.8 – Fifteenth harmonic amplitude data

A3: Bright specification

Harmonic	c	c1	c2	c3
1	221	214	209	203
2	199	190	197	185
3	173	152	156	150
4	154	137	137	130
5	139	117	117	115
6	133	107	107	91
7	117	101	101	59
8	99	84	84	54
9	92	61	61	40
10	88	69	69	
11	76	60	60	
12	70	49	49	
13	56	27	27	
14	43			
15	34			
16				

Table A.9 – Open Diapason harmonic amplitude data

Harmonic	c	c1	c2	c3
1	198	198	185	178
2	187	187	174	167
3	145	145	132	125
4	128	128	115	108
5	99	99	86	79
6	108	108	95	88
7	61	61	48	
8	84	84	71	
9	35	35	22	
10	63	63	50	

Table A.10 – Principal harmonic amplitude data

Harmonic	c	c1	c2	c3
1	191	186	184	182
2	164	172	142	137
3	127	118	114	101
4	129	123	106	96
5	102	101	46	62
6	86	90	38	49
7	86	80	24	
8	75	70	25	
9	63	54	18	
10	63	50		
11	43	48		
12	41	39		
13	45	38		
14	39	35		

Table A.11 – Twelfth harmonic amplitude data

Harmonic	c	c1	c2	c3
1	208	201	188	196
2	189	180	156	164
3	159	158	136	125
4	141	128	122	
5	137	105		
6	127	87		
7	98	79		
8	97	53		
9	75	29		
10	81	42		
11	68			
12	78			
13	60			
14	68			
15	55			
16	54			
17	30			
18	29			
19	28			
20	38			

Table A.12 – Fifteenth harmonic amplitude data

Appendix B: Blend test data from section 3.6

			Test 1			Test 2			Test 3		
Age	**Location**	**Ability**	Preference	Blend	Strength	Preference	Blend	Strength	Preference	Blend	Strength
old	Canada	Listener	1	4	2	8	8	8	2	2	5
young	USA	Excellent	6	6	6	8	10	5	4	2	8
young	USA	Good	8	8	7	3	3	5	8	9	8
young	USA	Good	11	11	11	1	1	11	11	11	11
old	USA	Moderate	3	2	5	10	10	7	6	6	6
old	USA	Excellent	3	3	3	9	9	9	4	4	4
young	UK	Good	4	4	9	3	8	3	6	7	6
young	USA	Good	11	1	11	1	11	1	11	1	11
young	UK	Moderate	7	5	7	4	8	4	4	4	7

Table B.1 – Data from subjects given a definition of blend

		Test 4			Test 5			Test 6		
Location	**Ability**	Preference	Blend	Strength	Preference	Blend	Strength	Preference	Blend	Strength
Canada	Listener	8	10	6	2	2	2	2	2	2
USA	Excellent	6	6	6	5	5	5	3	3	5
USA	Good	6	6	6	6	6	6	8	9	8
USA	Good	11	11	11	1	1	1	11	11	11
USA	Moderate	6	6	6	9	9	6	3	3	5
USA	Excellent	6	6	6	4	4	4	6	6	6
UK	Good	6	6	6	4	4	8	6	8	8
USA	Good	6	6	6	1	11	1	11	1	11
UK	Moderate	8	4	8	4	8	4	4	4	7

Table B.1 continued (with repetition of location and ability for clarity)

			Test 1			Test 2			Test 3		
Age	**Location**	**Ability**	Preference	Blend	Strength	Preference	Blend	Strength	Preference	Blend	Strength
old	Holland	Good	10	9	7	2	3	4	9	9	7
young	USA	Good	6	6	6	3	6	5	5	5	6
old	USA	Moderate	3	3	5	8	8	7	2	3	4
old	UK	Good	5	4	9	11	11	6	3	2	7
young	USA	Excellent	3	4	9	3	7	7	3	4	6
young	UK	Good	4	4	7	7	9	3	4	3	8

Table B.2 – Data from subjects not given a definition of blend

		Test 4			Test 5			Test 6		
Location	**Ability**	Preference	Blend	Strength	Preference	Blend	Strength	Preference	Blend	Strength
Holland	Good	4	4	4	4	4	5	9	9	7
USA	Good	6	6	6	4	6	6	7	6	6
USA	Moderate	5	5	6	4	4	6	2	2	5
UK	Good	7	7	5	6	7	5	1	1	8
USA	Excellent	6	6	6	8	8	8	4	4	4
UK	Good	7	5	6	7	7	4	4	3	8

Table B.2 continued (with repetition of location and ability for clarity)

Appendix C: Word gathering raw data from chapter four

Subjects' responses are presented here in raw form. Incorrect spellings, apostrophes and capitalisation have been corrected, but the grammar and other characteristics of the answers are unedited. Responses are numbered to allow precise referencing in the text, but individual subjects do not necessarily have the same numbering throughout all organs in a test, particularly as some subjects did not give a response for all organs.

C.1 Subject Responses to Organ 1, Test 1

Audio example: St. Columba's. Image displayed: St. Columba's.

1. Rich, mellow sound. Lots of body
2. Smooth, bland, pleasant, but not exciting. Not much "fire" to the pipes--sounds like heavily nicked pipes of the 1920s, leaning toward the English tradition. Needs some life; the room itself doesn't sound very reverberant, and there was a need for some higher harmonics or mixtures in the pipework.
3. Sounds kind of blah to me, nothing stands out, not a lot of character, just your basic organ tone. Not objectionable in any way but nothing to write home about either.
4. Light sounds, gently voiced, fairly clear sound, overall pleasing to the ear. A basic diapason chorus sound. Dead acoustic.
5. Nice full sounding diapason chorus, majestic, yet subdued, but with a solid fundamental tone
6. Sweet, calming, restful. It has the look and feel of a 19th century US instrument. Probably nice for congregational purposes.
7. It sounds like a 1950s Moller -- 8' and 4' Stops and a 2' with no mixtures or reeds. It might be a nineteenth century organ, also, it is probably a pre- 1910 or post 1940 organ.
8. Conventional Church organ - romantic voicing probably UK - 2 manuals? Great Diapason and Principal 8' + 4' designed for congregational singing. J W Walker 1950s?
9. Very clean principal chorus well scaled and finished for the dry acoustic in which it sings. The sound is modest in relationship to the visual impact of the facade and case. The congregation, I feel, is fortunate to have this organ leading their song with its cheerful plenum. Hearing the sound over the internet with less than wonderful sound

reproducing equipment I do believe the sounds could be produced by a fine quality electronic instrument as well as a pipe organ. It is not very articulate and as most electronics have too much artificial chiff it's probably a pipe organ.

10. The organ has a very full bass sound and strong high sound. The middle is weak.
11. First phrase of Old Hundredth was played on principal sounds at 8 and 4 foot pitch. I did not hear a pedal part.
12. A richer sounding pleno than I was expecting from the picture. Reverberation is pretty minimal; I would choose a more legato approach to the phrase played.
13. Basic English-style diapason chorus, reasonably balanced although 2' pitch is prominent in comparison to 4'. Room adds little reverberation.
14. Average sized parish organ, clear voicing, not muffled, not chiffy.
15. Breathy starting partial on every note - sounds as though very unresponsive action! Lacks warmth, 8ft component is probably very narrow-scaled (too narrow for a dry acoustic environment). 4ft element is bright and possibly a little strident - not well matched to the 8ft or the bass. Hard, dead, thin, forced sound.
16. Soft, doesn't sound overblown in any way. Reverberation (not after last note!) existent but not "blurring" to the sound - notes well resolved (but that probably has more to do with the architecture rather than the organ!).
17. Small parish church sound, principal sound.

C.2 Subject Responses to Organ 2, Test 1

Audio example: St. Chad's. Image displayed: St. Chad's.

1. Exciting, cutting with the cornet, Clean sound.
2. Very rich sound - lots of interesting harmonics, but not screechy as one might expect. Almost creamy, although that could be to do with the acoustic. Has a lot of character - it's an organ I'd like to hear more of.
3. Sounds a bit brighter, richer harmonics, but not much depth to the lower pitches; still sounds a bit British; I like the sound a bit more initially than organ one, but not sure it would wear well. Needs more foundation to it.
4. This organ seems to be in a room offering a much livelier acoustic than the previous one. Also, the registration used has a much more nasal quality, and sounds "thinner" and not so broad or tubby as the previous one. Nonetheless, the character is still very grand and majestic - probably due to the registration chosen.

5. Nice resonant building. Possibly a Catholic Church. Well sited in West gallery. 2 manuals + pedal. Great 8, 4 + swell 4,4,2 diapason to oboe. Effective for hymn accompaniment. Pleasant resonant sound.
6. This organ seems a little edgier (if that's a word), the sound isn't as plain, it could be said to have a little bite to it. It's a more interesting sound.
7. This sounds like a 1920s organ. There is a tierce either by itself or in a mixture. There are no reeds being played.
8. Nice full sound. The bass, middle, and top are strong. It's equally balanced.
9. A 16 foot Pedal gave depth to the bass line of this playing of the first phrase of Old Hundredth. Again principal or diapason tone at 8 and 4 feet predominated.
10. Very elegant sounding pleno and reeds. Reverberant room enhances greatly.
11. Pipe organ with a reedy clang and well-defined pedal. Good balancing of the plenum, articulate voicing allows listener to hear the rhythm clearly within the context of legato style articulation in a lightly resonate room.
12. Again, looks like a 19th century instrument. This one perhaps not US. Nice acoustic makes the organ sound better and yet this is not a rich sound like the first instrument. Too bad I can't go back for comparative purposes. This is not a warm sound and yet it is not unpleasant. First organ had perhaps a round diapason sound. This one more principal tone. This room really enhances the organ.
13. Much "lighter" in sound - more airy. Reverb much higher than for the first one - evidently in a building with more hard surfaces. Causes a little blurring of notes (compared with last one!). Much more "flutey".
14. Much more blending acoustics! The individual ranks bind together in an ensemble much better than in the first example; the sesquialtera has warmth enough to bind the other elements into one composite timbre rather than multiple sound ingredients fighting one another. Reedy, piquant, warm, full, lively. This organ sounds earlier (18th century) than the one depicted - lower wind pressures, gentler sweeter voicing.
15. Beautiful principal chorus, a little more "stringy," with prominent nazard, or perhaps a chorus reed. Room adds considerable "bloom" to the tone.
16. A driving tone with a metallic edge to it, well balanced upperwork,
17. Reed added to sound, some upperwork, more reverberation in the room, certainly a more exciting sound than Sound #1.

C.3 Subject Responses to Organ 3, Test 1

Audio example: Jack Lyons. Image displayed: Jack Lyons.

1. I assume that's not full organ. It's a flutey sound, breathy, without much foundation. Rather buzzy. Organ sounds smaller than it looks.
2. Sound has some edge to it because of voicing with limited nicking. Pleasant but not exciting.
3. This organ has a shimmer to it. The sound is not flat but has some life in the individual chords of the piece. It's more interesting still listen to, as there is more to hear than just the notes themselves. There is also an openness or hollowness to the sound (hey, you said the description could sound non-sensical)
4. It sounded like a bright modern classical style instrument - a college perhaps? I went to look at the RCM site to see if it was the new room 90 organ, and I don't think it's the Lyon's concert hall.
5. Oddly enough, this organ sounds very similar to one used to accompany some of the anthems on a CD that accompanied a catalogue from Morning Star Publishers here in the states. The sound is very "open" -- does not have the nasal qualities of the previous organ. At the same time, the tone quality is a little harsher too, in regards to the speech of the pipes. As such, the articulation is much clearer sounding
6. Sounds better, but rather fuzzy and not very articulate. Seems to have better reverberation, but not very brightly voiced.
7. This sounds like a modern tracker built in the 1950s or 1960s. There are no mixtures or reeds being used. The tone is clear and articulate.
8. Gorgeous case, apparently in a concert hall. Nice clean pleno with decent reverberation in the room. I'd like to hear the additional guts I suspect this instrument has.
9. Northern European sound. Strong treble could possibly have melody soloed out, as could the previous two examples. Quality of reproduction prevents me from knowing for sure. Tone is more flute like rather than presenting the bright full spectrum of harmonic development heard in traditional diapason sound.
10. Obviously a new instrument - could be as old as 1960s judging from the layout. This one has to be in a concert hall. A little buzzy. Something lacking here. Not enough fundamental perhaps. It's almost a bit anaemic. Disappointing. Perhaps there is no 8' playing. Harmonics too strong. Uninviting.

11. This organ has a much more piercing sound with the high pitches. There is also a breathy attack to the notes.
12. Higher-pitched bright stops were added. 2 foot perhaps to the 8 and 4 principals from the Great, plus the depth of 16 and 8 and perhaps even a 4-foot stop in the pedal.
13. More baroque-sounding principal chorus. Bright character. Fundamental (8') tone is a little thin, tending toward fluty. Works nicely with the moderate resonance of the room.
14. Sounds overblown - can hear air rushing through above the sound of the notes themselves. Seems to have a kind of tape noise sound in the background - effect of the organ on the recording? This wasn't pronounced on past recordings, though. Higher notes more piercing/harsh to the hear - not as pure.
15. I had my first organ lesson on this organ! Surprisingly unpleasant to listen to - maybe a factor of the .wav file format - neo-baroque organs tend to need high quality audio reproduction to do them justice. 8ft has excessive 3rd harmonic (quint) in relation to its fundamental; again lacks warmth and has brittle sound. 4ft is positively edgy - obviously voiced to sound as bright and projected as possible. (omission of staring partial on first chord may exacerbate this) comes across as abrasive and forced - would be tiring to listen to after a short while.
16. Chiffy attacks, baroque or neo-baroque style instrument, maybe with flexible winding? I think I'd quickly tire of the chiffiness.
17. not well balanced, weak fundamental, slightly "fluffy" whilst attempting to be metallic

C.4 Subject Responses to Organ 4, Test 1

Audio example: Heslington. Image displayed: Heslington.

1. Clean, cohesive principal chorus sound. Some degree of brilliance to the sound.
2. Good cohesive chorus. Otherwise fairly unremarkable - not too much character.
3. Very pleasant principal chorus on this organ. Very slight nasal quality present, apparently due to the way the principals were voiced -- the upper harmonics are noticeably present in greater quantities. The fundamental 8' tone isn't quite as prominent in this chorus, though -- the upper work seems to dominate the tone (I'm assuming that this is a 8'+4'+2' principals registration)
4. Well this example is played faster, I'm not sure if that will have an effect on what I say or not. I find this to be similar to the first sample although a little more

interesting/less homogenous. The high pitch in the last chord seems a little out of place. I'd really like to reflect and then listen again but that's probably against the rules...

5. This organ sounds like it could have been built in the 1950s or 1960s. Here again are no reeds or mixtures, this time no tierce. It has a clear tone, not forced, and is on low wind pressure.
6. Unusual colourful case layout. While modern almost has a deco look. Even American Southwest. Room acoustic a little disappointing. Once again somewhat lacking in warmth. Not as bad as example #3. Tone is clear. Maybe this is just 8,4 and 2. Brightness a little uneven. I was impressed with the warmth and smoothness of #1 but it probably would not be well suited for anything beyond congregational accompaniment. Certainly Bach would not be suited to it. #2 still relatively had a smoothness and evenness but with a little more brilliance. Maybe there's a little more fun to #2. The choice of registration may have hurt #3. Given that it appears to be a large instrument, it probably has great musical range, but still it was harsh and would probably be tiring after a short time. #4 is probably a decent compromise. It has some of the clarity of #2 and could in fact sound similar in a similar acoustic. Could #4 be a recycled instrument?
7. It doesn't sound that different to the other instruments - maybe my hearing is deteriorating with age! Again, it's well sited thus giving a clear sound. It looks as if it's a North European organ installed in the last 10 years or so. It would be nice to hear a longer sample - though I'm not sure what the purpose of your research is. Are we to be told?
8. Perhaps a mixture has been added to brighten the four-part harmony from the manuals even more. The pedal supports with the 16, 8 and 4 foot stops.
9. The pitch doesn't seem stable on the attacks. The middle sounds strong than the bass and the top.
10. Principal sound with mixture, light pedal. Clearly defined melody line. Articulation result more of detached playing than fine voicing. Acoustical response does not match the photo with its hard wooden ceiling, proportion of height and width and apparent hard surface on the walls. Makes an interesting study of how the visual element of the organ art effects what the listener thinks he is hearing. In this case, I find the organ to be more interesting visually than aurally. If I were to see a broken down electronic organ and hear the fabulous sounds of a French cathedral organ coming from it I

would have trouble enjoying the experience as much as if I were in the French cathedral listening to the same organ sound. This organ has a very bright sound, perhaps it would sound more balanced in a live hearing in the church.

11. Again, higher notes not so pronounced. Reverb not non-existent (existent!). Cannot hear the sound of rushing air through it over the sound of the notes themselves. Bit more "blocky" in sound than e.g. no. 2 -- that is, not as elegant.
12. Somewhat brighter principal chorus, rather neo-baroque in character. 8' tone is weak in comparison to upperwork. (Maybe because the 7-rank organ I usually play on Sundays has virtually no upperwork.)
13. A striking case with very pleasant pleno, upperwork crowns pleno very nicely. Probably my favourite of the instruments thus far.
14. I have played this organ relatively recently. Doesn't sound like it looks - probably because half the pipes were second-hand when they built it! Sound is not unpleasant - seems well matched to the acoustic space - not forced, but projected. 2ft stop is louder than ideal - it stands out from an otherwise warm and well-blended texture. 8ft has sufficient fundamental to support and enough partial development to lend clarity. 4ft blends well. Clear, pleasant, interesting but not aggressive, solid but not ponderous.
15. “simple” upperwork, limited driving tone to it, very English parish church, rather “held back”
16. Bright sound, upperwork used, recorded farther away than was Organ #3. Relatively thin chorus sound.

C.5 Subject Responses to Organ 1, Test 2

Audio example: St. Columba's. Image displayed: St. Chad's.

1. A bit dull - sounds like a small carpeted room.
2. It has a very pleasant sound, a moderately bright chorus with just the littlest bit of a hint of thickness to the sound, with an appropriately firm pedal tone. If more than one register is being used, they do seem to blend rather nicely. The room does not come across as particularly live.
3. Sounds just like most church organs, - adequate for the job, but nothing special about it. Not an organ that I would go very far to hear.
4. This organ sounds very typical (to me) of organs built in America in the late 1800's by builders like Hook or Felgemacker. The dominant voice I hear is what seems to be a 2' principal but somewhat of a 'horn diapason' quality rather than the Germanic principals used by contemporary American builders. The 8' principal (if there is one, it sound more like a flute-ish combination of stops rather than a principal chorus) sound too small scaled to my ears...not particularly robust. Perhaps it is a very harmonic 8' Gamba, which could produce a similar sound.
5. A thick and "woolly" sound. Mostly fundamental tone, few upper partials, probably wide windways and higher cut ups.
6. A crisp clear sound with good articulation in the upperwork. The building does not sound terribly live but there is some reverberation present which is helpful. If this is the principal chorus, it seems a bit dominated by the upperwork but that could be the location of mikes etc.
7. "English" (meaning UK) diapason chorus - rather reedy/stringy in quality, Willis?
8. A nice, not to shrilly, not too dull organ sound. 8' and 4' used. Little reverberation, intimate acoustic.
9. Diapason organ tone 8' fundamental plus one prominent partial at 2', no trem, very little reverb, quiet, firm sound. Not exceptional.
10. Bright, singing, a little stringy, a bit lazy of speech
11. It's a nice, tranquillising sound: Not too dull, not too shrill. People in worship would like it. The acoustic is a little "dead" (practically no reverberation).
12. This organ projects well into the sacred space it serves. It accompanies congregational singing excellently because of its clarity, its marriage to a wonderful acoustic environment and its warm, yet clear, sparkling voicing.

13. Rich sounding Diapason, resonant, room is excellent setting for organ sound.
14. Clear, full-bodied tone, plain, speaks clearly and cleanly, some stringiness, no harshness or excessive odd overtones.
15. "Thin" -- the top and bottom seem most prominent. To me it's not top-heavy or bottom-heavy, but rather "mid-range light."
16. Agreeable, rather smooth chorus sound. Good melodic projection. Assuming this is a diapason chorus would like to hear a brighter balance with some 'sparkle' (though not 'steeliness') at the upper end
17. Full round, woodwind.
18. Resonant, vibrant, solid sound. Commanding

C.6 Subject Responses to Organ 2, Test 2

Audio example: St. Chad's. Image displayed: Jack Lyons.

1. Strident, very nice sounding - surprisingly little chiff. Nice!
2. Still a woolly sound to me. I detect quite a lot of tierce in the ensemble. on computer speakers it's really difficult to sort out things as I would at the voicing machine in the shop. This still has a great deal of fundamental tone, and uppers are not very prominent to my ear. Not really a sound that I care much for, not very interesting.
3. Clearly a much more reverberant space than the first sound! And obviously more than just a single stop has been drawn. A prominent quint can be heard in the chorus; from the tonal quality it sounds like a 12th rather than a mixture being drawn. There is the chance, too, that the melody is being soloed out; there's also a bit of a 'third' or Tierce quality standing out in a few spots. If there is a pedal voice being played, it's not very discernible if at all -- a little bit of the end of the example sounds as if there's a very light 16 in the bass.
4. This organ sounds more like the product of contemporary organ builders. The principal chorus is not totally unlike the first sample, but the ensemble sounds to me like it has a tierce-mixture. Again I find that the principal is somewhat lacking in the kind of "oomph" or foundational tone that I prefer. The mixture is much more in evidence and I think that the organ is tuned in a non-equal tempering. The plenum has a very "reed-y" quality typical of the north German/Dutch baroque organs.

5. There is a lot more quality in the sound of this organ. I am afraid that a few bars of a hymn do not do it justice, - I am sure that it sounds better in real life than it does here. I like it, - where is it?
6. G D&B Organ (in the picture) and probably in the recording. UK Baroque revival sound from the late 1950s and early 1960s. Tries to be more "continental"/"classical": harmonics 5th, 12th and 17th more prominent than Organ 1.
7. Somewhat strident and "crunchy" sound with a bit too much low quint tone. Inarticulate.
8. The overall effect is a bit strident for my taste, but the brilliance of the upperwork certainly cuts through any muddy effect caused by the liveness of the building. Articulation is excellent and the harmonic development of the chorus first-class.
9. This organ and the room it serves seem to project a somewhat 'quinty' sound, a little more bright and cerebral than Organ #1. What a fine acoustic placement!
10. The sound is silvery and exciting... we call it a "neo-baroque" sound; ideal for playing Bach or Handel, for example. The acoustic has a fine reverberation which brings life to the organ sound.
11. Nice chorus 8 4 2. 2 is prominent, maybe too much - nice reverb - no punch.
12. Brighter, livelier, more (?) interesting chorus sound than organ #1 good melodic projection although the extra upperwork did not seem to do much to aid the projection of the melodic line. Slightly reedy quality – maybe inclusion of a quiet reed or a tierce in the mixture(s)? The overall timbre is perhaps somewhat 'lean' for the musical example (hymn-tune) played.
13. Full sound, but not entirely well blended due to prominent mutations (screaming mixture, perhaps?).
14. Thinner sounding than organ 1, lots of odd harmonics (a Tierce?), much geigen stringiness, not as cleanly speaking.
15. Bright, less dynamic, smaller sounding Diapasons. Resonant. Less commanding than first sample.
16. Rich & complex but seems top heavy. Maybe middle frequencies are missed.

C.7 Subject Responses to Organ 3, Test 2

Audio example: Jack Lyons. Image displayed: Heslington.

1. Chiffy, flutey. A disappointment.
2. Much better clear singing sound of the Principal with this. Plenty of 4' over tones, but still quite full in 8' fundamental tone. This sound is more in keeping with our work. I'm assuming that these are all organs in the U.K., following an "English" tradition of building and voicing?
3. This sounds like it's all on manuals, with the melody soloed on a registration including upperwork not contained in the accompanying voices, something like 8' and 4' flutes for the accompaniment, with 8' and 2' or 8' 4' 2' flutes for the melody. Nice acoustic, too!
4. This organ sounds like the ensemble used was topped with a 2' blockflute and a much more 'flute' like registration. More gentle to the ears than sample no 2 and more clean sounding. The principal at the bottom seems a bit 'rounder' that the sample of the number 2 organ, but the flute stop "on top" dominates. A very pleasant sound.
5. Seems to be one of those organs that we sometimes come across, all top and no bottom! I am not so sure that it is the organ, it may be the registration, - not being used well.
6. A mellow sound, much more reserved than the first two organs in this survey. Yet the sound is mellow without any trace of muddiness, retaining good definition and a lovely blend of the harmonics of the various components of the chorus. Quite a romantic sound.
7. UK/English instrument from 1960's / early 1970s. Light pressure, open foot voicing (minimal nicking). Non-unison harmonics less prominent than organ 2. Light and "open" sound. 8ft probably of the flute variety rather than a diapason/principal.
8. Although this sound is neither shrilly nor neo-baroque, and no mutation or high pitch stops are used, it has a piercing compound which annoys me and makes me nervous and uncomfortable. The amount of reverberation is very good: not too much, not too little.
9. Flutey in general, but with a bit of scratchy edge to the sound. 4' a bit overbearing.
10. Something funny sounding - higher partials - too much hollow sound, grating on nerves, nasal sound, some wobblies in 2' - don't hear much reverb

11. The sound of the 8' Principal of Manual I is clear, distinct, yet warm and comfy to listen to. The room has just the right amount of projection of the organ's sounds distinctly, clearly and without the acoustic interfering with the clarity of each and every single note.
12. Bright, resonant, small scale Diapason, sweet sounding
13. Diapason is much smaller and speaks with less authority that first organ sampled. Obviously a big room as evidenced by massive casework. It was a sweet sound.
14. Full, blended sound. A little "breathy" and "chiffy."
15. Less fundamental, more even harmonics -- especially 2nd harmonic, maybe more tin in the pipes, lacks weight and character.
16. Good, bright chorus, not sparkling (no mixtures in use?) but clear and 'easy on the ear'. Well-defined melodic line. Either diapason(s) with a fairly harmonic content or a flute component as the basis?
17. Pleasant flute sound. As played the bass is thin. Other registrations are available

C.8 Subject Responses to Organ 4, Test 2

Audio example: Heslington. Image displayed: St. Columba's.

1. Nice cohesive chorus - not overly bright.
2. This organ, like organ number one has a pronounced 2' register (or a very harmonicky Gambe) but I think that the 8' principal has more foundational tone than organ number one did. I also think that this organ is a new organ in an old case. As I re-listened to the sample, it sounded like it has a more fully developed principal chorus than organ number one and approaches the chorus development of sample number 2. The organ is bright without being too aggressive.
3. Now this one has potential! I like the balance, and the organ seems to have a singing quality that lies beneath the surface of the sound. Where are these organs, - might I hazard a guess that they may all be in York?
4. Even before listening, my suspicion is this is going to be a 'dry' sounding room. (now listening) Well, looks are deceiving! There's a bit of liveness to the room. The instrument has a nice bright tone, a bit thin, not as much body as the first. There doesn't seem to be any mutation work involved -- if there is, it's very very mild, which in a lot of cases is a good thing. Not much prominent pedal (16') voice here either.

Interesting that none of the examples seemed to have a typical chorus mixture involved.

5. Muddy weak fundamental tone, over shadowed by upper work. Not very even chorus. I least like this sound of all – even the first woolly organ.
6. Not the sound I expected from the photograph! I expected an instrument with that was rather duller in tone with a stronger unison and less reverberation. This organ sounds quite bright, the chorus seems to be built on an open diapason / principal and blends well together. The sound is clear.
7. This is my least favourite of the four. The articulation seems sloppy and the chorus seems less of a unified whole than I would like to hear. The upperwork seems to sit independently upon an unrelated 8' rank. The building seems quite dead which does not help the blend. For me, an unsatisfying sound.
8. Gentle, dignified but somewhat thin and non-blending in the treble. Inarticulate but not lazy.
9. Bright, clear, projects down the centre of the nave in a gentle but commanding way. It is ideal for leading a congregation in worship and in singing. Bravo to the builder!
10. Better balanced than 2 or 3, enough stringiness to give some bite but not too much to thin out the sound, could have more fundamental to make the best organ sound in this group.
11. I would describe it as very like organ 2. The curious thing is: It's not the sound I would expect from this organ. In the first 3 examples the sound matched with the organ's facade style. The echo causes a little "smearing" in this example.
12. Nice sound, but I hear beats at the end chord. Is the 2' in tune? Little reverb.
13. Fairly well blended sound -- highest pitch has a little bit of prominence, almost as if not in-sync with the other pitches.
14. Bright, not as resonant as previous samples, probably due to smaller room. Not as commanding as organ #1
15. Bright sounding, smaller room and less resonance. Obviously a good service organ. Still not as much presence as organ #1
16. Robust, full chorus sound with a distinct 'sizzle' on top, particularly noticeable on the last chord of the excerpt. Would become fatiguing to the ear if played chordally for any length of time without the aural relief afforded by intervals with a change in dynamic level, timbre or both. A 'busy' sound which takes attention away from the melodic line to some degree.

17. Rather plain sound which is not unpleasant though could be 'twee'

C.9 Subject Responses to Organ 1, Test 3

Audio example: St. Columba's. Image displayed: Heslington.

1. Sounds OK
2. To me the sound is an 8' Diapason played alone.
3. warm, supportive chorus for congregational singing; pleasing organ tone; no eccentric voices
4. Nice... the right hand sounds nice and "flute-ish." The left-hand (tenor) range sounds a little reedy or stringy. The reverb in the room is nicely dry to hear the pipes clearly, but not dry enough to make it unpleasant. The pipes sound quickly, and I don't hear any chiff. Holland?
5. It's sounds rather dead in the location, whereever it might be. However, the notes are clear.
6. Good, normal sound, nothing unusual.
7. Heavy, lacking articulation. Bass too loud. Adequate harmonic development but needs regulation to match the acoustical environment.
8. A bright and relatively immediate sound. Not a very large acoustic space. Some reflected sound allows just enough legato in the sound so as to produce a singing quality, but I suspect that when the building is full of people it would sound quite dead. The Principal quality is more continental than English - a bit brighter and not quite so tubby.
9. Clear. Clean. Mellow.
10. This organ strikes me as a very competent, but not particularly exciting organ. There is a preponderance of 8' and 4' tone, and an absence of brightness that might be associated with a mutation or mixture. Voicing is clear and consistent with the acoustics in the space.
11. Fairly mellow tone. European tone. Diapasons very strong, sound like just diapasons.
12. First, I was struck by the sudden "burst" of the initial note, which made me suspect it might be electronic -- but which might also be a by-product of editing the file for size. Next, it seems in good tune, and the registration balanced -- 'though my tinnitus may be deceiving me. The "space" seems a bit small perhaps -- I assume that to be the

reason for the rest after the first note (so we could hear it). I would happily listen to it on a "regular basis" if it were in my church -- assuming that other registrations also exist!

13. Mellow, smooth. English diapason sound. Small. Close.
14. The overall sound is thin and reedy. If a Principal/Diapason of 8' pitch is included in the registration, its scale is quite small. A 2-2/3' pitch component is strongly heard, suggesting that there is either an independent voice of this pitch in the registration, or a voice with a strong twelfth component, such as a Quintadena 8'. Since these examples are being played through my computer's speakers, I am not sure how faithful they are to your original recordings.
15. Good acoustics. The organ lacks a bottom - no 16' pipes are visible or heard.

C.10 Subject Responses to Organ 2, Test 3

Audio example: St. Chad's. Image displayed: St. Columba's.

1. Either this church is way bigger than it looks, or that's not the right picture!
2. Yuck. The acoustics are a bit to wet for me--sounds a little slurred. Sluggish. The ranks don't seem to blend together. I hear little ranks popping out in different places.
3. This organ is a bit more lively that the previous organ. It also sounds as if it's in better tune than the previous organ.
4. 8' Diapason with added stops sounds to me like possibly 4' and 2 principals. This is a fuller and brighter sound than sound no. 1.
5. Bright, robust, but not overbearing; good for leading singing; a bit thin in overall effect
6. Bright, thin, still pleasant to my ear.
7. Good acoustics, a lot of reverb.
8. Wonderful. Balanced, cohesive chorus. Great hymn leader. I think I'll sing!
9. Reedy, bright smooth. Less close than No 1. As if in medium size church.
10. This organ has a bright, engaging quality. The registration includes voices with higher order partials which complement an acoustically "live" space.
11. Brighter. Pedals sound slow speaking toward end of clip. Nice variety of stops sounding.
12. Far less precise in attack ("smearing" of sound? -- If a picture, I'd say badly focused). I found it bothersome to listen to, although I could hear "all the notes" -- the

harmonies etc., and the registration seemed the same. I don't "think" the longer reverb had anything to do with it -- I couldn't get a sense of the "space" of the room in listening. I'd still have to say that I didn't enjoy it.

13. One might assume from its visual appearance that the sound of this organ would be characteristic of the early part of the 20th century. That it is not suggests that the organ has been substantially rebuilt, or a new instrument installed within an old case. The sound of this instrument is even thinner and reedier than the first example. The twelfth pitch is particularly prominent in the registration. The room is more reverberant than the first example. One might assume that the registration contains pitches 8.4.2-2/3.2
14. Excellent reeds in a lively acoustic.

C.11 Subject Responses to Organ 3, Test 3

Audio example: Jack Lyons. Image displayed: St. Chad's.

1. Flutier and less fundamental than would be expected, but nice.
2. Again an 8' Diapason with possibly a 4' Flute is a softer warmer sound which may also be due to scaling. Reminds me of 4' Harmonic Flute
3. Pleasing, warm sound, yet a little timid for assisting congregational singing
4. The pipes seem to speak from the top town...almost like an arpeggio. The bass note is always a little behind the soprano. The soprano range is much more pungent than the tenor (left hand). Somewhat like a church choir that has a really strong soprano, weak altos and tenors, and a bass who can't keep up with the rest of the people.
5. Good. A little flutey, but nice. Melody leads a bit too much. Scaling too broad in the middle range? Slightly smaller scales in the 4' would have been beneficial. Good overall.
6. This organ doesn't sound quite as dead as the first organ but it's not as lively as the second organ. Also certain pipes are sounding more loudly in this organ than what I noticed in the first or second organ.
7. Best yet. Distinct.
8. Baroque sounding pipework, good acoustics.
9. This organ has a dominant 4' open flute which sounds like it has been combined with a softly-voiced 8' gedeckt. The space is lively and reverberant, though I get the

impression that the organ probably doesn't do particularly well when all pews are occupied.

10. About halfway between samples #1 and #2. I've never voiced pipes (something I'd like to learn), but the mixture sounded "rough" to me, not a stop I'd particularly choose. Of course, I "often" choose available MIDI "stops" that I don't particularly appreciate, because "they're there", and what's available. Oddly (to me), the reverb time seemed longer than the #2 sample -- perhaps like a "long, thin hall" (though not all *that* long).
11. Flutey, spiky, bright. Chamber organ sound in church acoustic.
12. Nondescript 8' organ in medium to low level acoustics.
13. The sound of this instrument is less jangly than the previous two, largely because the twelfth component is reduced. The curiously open appearance of the case suggests that a much smaller organ than the original is now installed within it, or that speakers for an electronic organ reside within the case. (The sound quality of the recordings makes it difficult to tell whether the instrument contains pipes or is an electronic.)

C.12 Subject Responses to Organ 4, Test 3

Audio example: Heslington. Image displayed: Jack Lyons.

1. A somewhat fuller and more integrated (i.e. blended) sound than previous examples.
2. Nice. Maybe out of tune? I hear "beats" on the last chord of the clip. The first note of the soprano seems to speak louder than the rest of the notes (the G-natural perhaps, if in the key of G-major). I really hear the words "Praise" and "Bless...(ings)" in the doxology. The "flow" syllable seems to almost disappear (the B-natural). Massachusetts, USA?
3. Bright, solid, prominent 15th on last note.
4. Again using an 8' and 4' stop principal tones these not as warm a sound as #3 narrower scales more North German in nature.
5. Articulate, strong, a bit of an edge.
6. Great. Bold but not intrusive. Good voicing and finishing. Good scaling. Who did this organ? Thanks be to God we have such builders!
7. Bright but thin.
8. This piece was played a bit faster than the previous pieces. Brighter, but still a bit on the dead side.

9. Majestic sound, rich sounding.
10. The sound of this organ is balanced and bright. Unfortunately, it is located in a space which is a bit dry for my taste. Nevertheless, higher order harmonics are pleasantly supported.
11. Fluty, fuller than No 3. Bright. Kingston Parish Church?
12. This sample (#4) is the first that's given me trouble in hearing "all the notes" in the harmony. Perhaps they were tuned a little too accurately? -- And were a too tight match for each other? -- Or maybe, just my ears. Back to a "shorter" reverb time. I'm afraid I've "flunked" the quiz -- sorry. Sample #2 almost gave me a headache, if it's any consolation.
13. Well balanced principals - clearly American.

Appendix D: Frequency of word occurrence

These are the words gathered from subject responses presented in appendix C and described in chapter four. "Inclusive" includes repeated use by the same listener, whereas "exclusive" does not.

Adjectives	Word length	Frequency (inclusive)	Frequency (exclusive)
Abrasive	8	1	1
Aggressive	10	2	2
Agreeable	9	1	1
Anaemic	7	1	1
Articulate	10	12	7
Authoritative	13	1	1
Balanced	8	12	11
neo Baroque	11	7	4
Bland	5	1	1
Blended	7	9	7
Body	4	2	2
Bold	4	1	1
Breathy	7	4	4
Bright	6	41	26
Brilliant	9	2	2
Brittle	7	1	1
Broad	5	1	1
Busy	4	1	1
Buzzy	5	2	2
Calming	7	1	1
Cerebral	8	1	1
Character	9	4	3
Chiffy	6	7	6
Clean	5	8	6
Clear	5	19	15
Close	5	2	1
Cohesive	8	4	3
Comfortable	11	2	1
Commanding	10	4	2
Competent	9	1	1
Complex	7	1	1
Creamy	6	1	1
Crisp	5	1	1
Crunchy	7	1	1
Cutting	7	1	1
Dead	4	4	2
Dignified	9	1	1
Distinct	8	2	2
Driving	7	2	1

Table D1 – Adjective frequency of occurrence data (continues on next page)

Adjectives	Word length	Frequency (inclusive)	Frequency (exclusive)
Dull	4	3	3
Dynamic	7	1	1
Edgy	4	4	4
Elegant	7	2	2
Engaging	8	1	1
Exciting	8	6	5
Even	4	1	1
Firm	4	2	2
Flat	4	1	1
Fluffy	6	1	1
Flutey	6	12	10
Forced	6	3	1
Full	4	11	9
Full-bodied	11	1	1
Fuzzy	5	1	1
Gentle	6	5	5
Grand	5	1	1
Grating	7	1	1
Hard	4	1	1
Harsh	5	3	3
Heavy	5	2	2
Hollow	6	2	2
Homogenous	10	1	1
Immediate	9	1	1
Integrated	10	1	1
Interesting	11	7	7
Intrusive	9	1	1
Jangly	6	1	1
Lazy	4	2	1
Lean	4	1	1
Light	5	7	7
Lively	6	4	3
Majestic	8	3	2
Mellow	6	5	5
Metallic	8	2	1
Mild	4	1	1
Muddy	5	2	2
Muffled	7	1	1
Nasal	5	2	2
Nice	4	14	11
Open	4	3	3
Overbearing	11	2	2
Piercing	8	4	3
Piquant	7	1	1
Plain	5	3	3
Pleasant	8	14	11
Punchy	6	1	1

Table D1 – Adjective frequency of occurrence data (continues on next page)

Adjectives	Word length	Frequency (inclusive)	Frequency (exclusive)
Pungent	7	1	1
Pure	4	1	1
Quiet	5	1	1
Reedy	5	8	7
Resonant	8	3	3
Restful	7	1	1
Rich	4	8	7
Robust	6	3	3
Romantic	8	2	2
Rough	5	1	1
Round	5	3	3
Scratchy	8	1	1
Screechy	8	1	1
Shimmery	8	1	1
Shrill	6	4	1
Singing	7	4	4
Silvery	7	2	2
Sizzly	6	1	1
Sluggish	8	1	1
Small	5	2	2
Smooth	6	3	3
Soft	4	3	3
Solid	5	4	4
Sparkling	9	3	2
Spiky	5	1	1
Steely	6	1	1
Strident	8	4	4
Stringy	7	7	5
Strong	6	7	5
Subdued	7	1	1
Sweet	5	4	4
Thick	5	2	2
Thin	4	14	12
Timid	5	1	1
Tranquil	8	1	1
Tubby	5	2	2
Twee	4	1	1
Uninviting	10	1	1
Vibrant	7	1	1
Warm	4	9	5
Weak	4	4	4
Weighty	7	1	1
Woolly	6	2	1

Table D1 – Adjective frequency of occurrence data (continued from previous pages)

Appendix E: Details of organs used

This appendix includes brief descriptions and specifications of the organs used to enable readers to put the stops and organs chosen in context. Pictures of the four organs as used in the adjective gathering experiment are also presented. All organs are in the United Kingdom. Specifications are taken from the National Pipe Organ Register (http://lehuray2.csi.cam.ac.uk/) with corrections made by the author.

E.1 Doncaster Parish Church

Doncaster Parish Church is the oldest, largest and most well-known of these six instruments. It was built in 1862 by the German firm of Schulze, and as a large early example of their work was highly influential on English organ-builders of the time. It consists of five manuals and pedals, with the Solo organ originally borrowed from the Swell. In 1910 Norman and Beard added an independent Solo organ and made minor changes to the stops, but tonally the organ is essentially as Schulze left it. The church is the size of a modest cathedral, although the reverberation is not excessive given its size.

Pedal				**Choir**	
Sub Principal	32	Fifteenth Bass	4	Lieblich Bourdon	16
Major Bass	16	Tierce	3 1/5	Geigen Principal	8
Principal Bass	16	Mixture	II	Viol de Gamba	8
Open Bass	16	Cymbal	II	Flauto Gamba	8
Violone	16	Contra Posaune	32	Gemshorn	8
Sub Bass	16	Posaune	16	Salicional	8
Major Bass	8	Bombarde	16	Flauto Traverso	8
Flute Bass	8	Contra Fagotto	16	Lieblich Gedact	8
Violoncello	8	Trumpet	8	Flauto Traverso	4
Octave Bass	8	Horn	8	Lieblich Flute	4
Quint Bass	10 2/3	Fagotto	8	Geigen Principal	4
Great Tierce	6 2/5	Clarion	4	Quintaton	4
Quint	5 1/3			Flautina	2

Great

Sub Bass (TC)	32
Double Open Diapason	16
Bourdon	16
Open Diapason No.1	8
Open Diapason No.2	8
Stopped Diapason	8
Hohl Flute	8
Stopped Flute	4
Principal	4
Gemshorn	4
Quint	5 1/3
Twelfth	2 2/3
Fifteenth	2
Mixture	V
Cymbal	III-V
Cornet	IV
Double Trumpet	16
Posaune	8
Trumpet	8
Clarion	4

Swell

Bourdon	16
Open Diapason	8
Terpodian	8
Echo Gamba	8
Voix Celeste (TC)	8
Harmonic Flute	8
Rohr Flute	8
Harmonic Flute	4
Stopped Flute	4
Principal	4
Viol d'Amour	4
Mixture	V
Scharf	III
Cornet	IV
Double Bassoon	16
Trumpet	8
Horn	8
Hautboy	8
Clarion	4
Vox Humana	8
Tremulant	

Solo

String Gamba	8
Harmonic Claribel Flute	8
Concert Flute	4
Clarinet	8
Orchestral Oboe	8
Tremulant	
Tuba	8

Echo

Tibia Major	16
Harmonica	8
Vox Angelica	8
Flauto Amabile	8
Flauto Traverso	8
Celestina	4
Flauto Dolcissimo	4
Harmonic Aetheria	II

E.2 Heslington Parish Church, near York

Heslington Parish Church is a typical two manual instrument of moderate size, speaking into a moderately reverberant acoustic. Originally it was a Forster and Andrews instrument of 1888 in a Methodist church. It was tonally altered when it was moved to Heslington, gaining mixtures and losing a second 16' on the pedal. The Great is exposed in the body of the church, but the Swell (not used in this thesis) is buried in the tower. The case dates from its 1974 installation in the church.

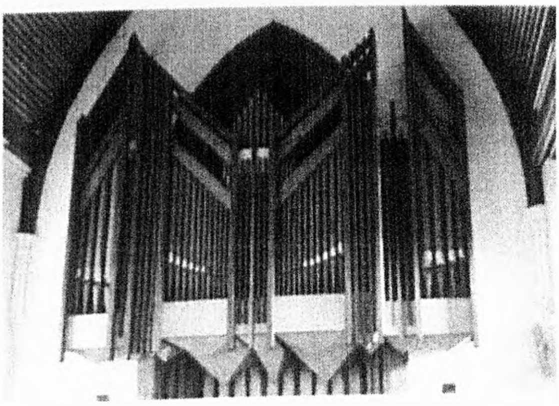


Figure E.1 – Heslington Parish Church organ

Pedal		**Great**		**Swell**	
Bourdon	16	Open Diapason	8	Stopped Diapason	8
Quint	10 2/3	Stopped Diapason	8	Salicional	8
Principal	8	Principal	4	Voix Celeste (TC)	8
Bass Flute	8	Stopped Flute	4	Gemshorn	4
Fifteenth	4	Twelfth	2 2/3	Piccolo	2
		Fifteenth	2	Larigot	1 1/3
		Mixture 19.22.26	III	Mixture 15.19.22	III
		Trumpet	8	Cornopean	8
		Tremulant		Oboe	8

E.3 Jack Lyons Concert Hall, University of York

The Jack Lyons instrument is the most recent pipe organ used, dating from 1969. It is in neo-Baroque style, built by the British firm of Grant, Degens and Bradbeer. It speaks into a moderately live concert hall. It was altered in 1983 by Walkers, who enclosed the Oberwerk, added a chorus reed in place of a mutation, and made other minor alterations in the interests of greater flexibility.

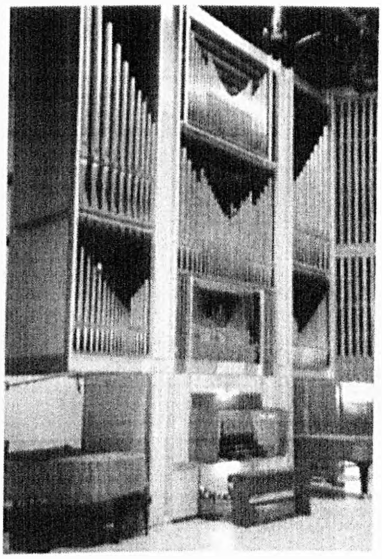


Figure E.2 – Jack Lyons organ

Pedal

Subbass	16
Octave	8
Rohrpfeife	8
Gemshorn	4
Bass Kornett	IV
12.17.b21.23	
Mixture	VI
Fagot	16
Rohrschalmey	8

Oberwerk

Holtzgedackt	8
Weidenpfeife	8
Principal	4
Rohr Flöte	4
Spitz Principal	2
Sesquialtera 12.17	II
Scharff 22.26.29	III-IV
Regal	16
Hautbois	8
Tremulant	

Hauptwerk

Quintadena	16
Principal	8
Spitzflöte	8
Octave	4
Rohrquinte	2 2/3
Flachflöte	2
Mixtur	V
15.19.22.26.29	
Trompete	8
Tremulant	

Brustwerk

Gedackt	8
Spitzgedackt	4
Principal	2
Nazat	1 1/3
Zimbel 29.33.36	III
Krummhorn	8
Tremulant	

E.4 Port Sunlight URC, Merseyside

Figure E.3 – Port Sunlight organ

The Port Sunlight organ is a large four manual instrument from 1904, speaking into a medium sized but not overly reverberant acoustic. It was built by Henry Willis II, and is essentially unaltered. Despite the presence of mixtures, the upperwork of this organ is more restrained than any of the other organs.

Pedal

Open Diapason	16
Violone	16
Bourdon	16
Bass Flute	8
Ophicleide	16
Posaune	8

Choir

Open Diapason	8
Dulciana	8
Gamba	8
Hohl Flute	8
Wald Flute	4
Piccolo	2
Clarinet	8

Great

Double Diapason	16
Open Diapason No.1	8
Open Diapason No.2	8
Harmonic Flute	8
Clarabella	8
Principal	4
Flute	4
Twelfth	2 2/3
Fifteenth	2
Mixture	III
Trumpet	8
Clarion	4

Solo

Flute	8
Orchestral Oboe	8
Vox Humana	8
Tuba	8
Tremulant (Vox)	

Swell

Lieblich Bourdon	16
Open Diapason	8
Lieblich Gedackt	8
Salcional	8
Voix Celeste	8
Gemshorn	4
Lieblich Flote	4
Mixture	IV
Contra Fagotto	16
Cornopean	8
Oboe	8
Clarion	4
Tremulant	

E.5 St Chad's Parish Church, York

St Chad's is a large two manual instrument dating from 1891 but rather old-fashioned for that time. It speaks into a moderate room that is very reverberant for its size, although it was originally built for a Lancashire Methodist church by a small Manchester builder. Unusual for its date are the mixtures incorporating a seventeenth component.

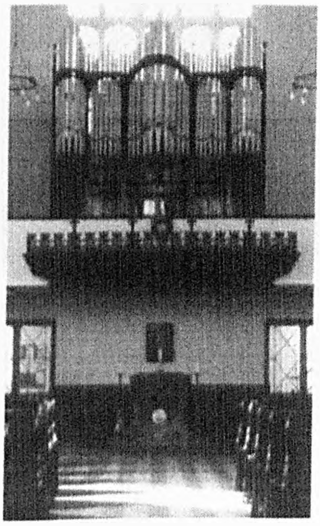


Figure E.4 – St Chad's organ

Great

Double Open Diapason	16
Open Diapason	8
St. Diap. & Clarabella	8
Dulciana	8
Viol de Gamba	8
Principal	4
Harmonic Flute	4
Fifteenth	2
Mixture (15.17.19)	III
Trumpet	8
Clarionet	8

Swell

Lieblich Bourdon	16
Open Diapason	8
Lieblich Gedact	8
Salcional	8
Voix Celeste (TC)	8
Gemshorn	4
Harmonic Piccolo	2
Mixture (17.19.22)	III
Cornopean	8
Oboe	8
Tremulant	

Pedal

Open Diapason	16
Bourdon	16
Bass Flute	8

E.6 St Columba's URC, York

St Columba's is a small two manual Lewis & Co. instrument from 1907, speaking into a relatively dead acoustic.

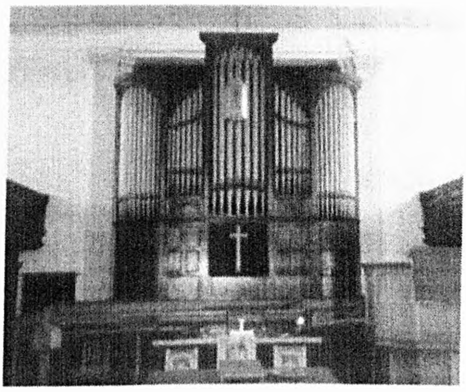


Figure E.5 – St Columba's organ

Great		**Swell**		**Pedal**	
Lieblich Bourdon	16	Open Diapason	8	Great Bass	16
Large Open Diapason	8	Stopped Diapason	8	Sub Bass	16
Wald Flute	8	Echo Gamba	8	Flute Bass	8
Dulciana	8	Vox Angelica (TC)	8		
Octave	4	Octave	4		
Harmonic Flute	4	Super Octave	2		
		Contra Fagotto	16		
		Oboe	8		
		Tremulant			

Appendix F: subject data for elapsed time experiment

Table F.1 presents the subject data on which analyses of subjects' ability to recall combinations of sound and images were conducted in section 4.5.6. Subjects are numbered to preserve anonymity, and the final column presents the number of correct answers. A correct answer here is one in which the combination between image and sound presented to subjects was later recalled.

Subject number	Test taken	Recognised any organ	Hours between tests	Minutes (italic) or days between tests	Correct answers
1	1	Yes	0.017	*1*	4
2	3	No	0.017	*1*	1
3	2	No	0.033	*2*	4
4	3	No	0.033	*2*	4
5	3	No	0.033	*2*	4
6	3	No	0.033	*2*	2
7	2	No	0.050	*3*	4
8	2	No	0.067	*4*	4
9	2	No	0.067	*4*	4
10	3	No	0.067	*4*	4
11	1	Yes	0.083	*5*	4
12	2	Yes	0.083	*5*	4
13	3	No	0.083	*5*	2
14	3	Yes	0.100	*6*	2
15	3	Yes	0.100	*6*	3
16	3	Yes	0.133	*8*	0
17	3	No	0.800	0.03	4
18	2	Yes	12.15	0.51	4
19	1	Yes	169.00	7.04	4
20	1	Yes	392.75	16.36	4
21	1	No	529.00	22.04	4
22	2	No	536.50	22.35	1
23	1	No	541.00	22.54	1
24	1	Yes	544.00	22.67	2
25	2	No	622.00	25.92	2
26	1	Yes	645.85	26.91	0
27	1	No	1004.00	41.83	4
28	2	Yes	1373.00	57.21	2

Table F.1 - Subject data for elapsed time study

Appendix G: Organ acoustic analysis data

This appendix includes the individual harmonic amplitude data for the six steady-state analyses of organs studied in chapter five. The acoustic signatures for these ensembles can be found in figures 5.1 to 5.6 of chapter five. All harmonic amplitude figures are in dB above the noise floor.

G.1 Analysis data for Doncaster Parish Church

Harmonic	C	F#	c	f#	c1	f#1	c2	f#2	c3	f#3
1	31	30	43	43	40	32	35	42	27	41
2	40	19	37	45	49	45	39	42	45	37
3	43	40	39	36	39	37	44	38	35	35
4	40	39	46	36	22	42	37	34	34	27
5	14	21	17	24	18	20	28	18	20	11
6	44	32	41	42	41	44	35	37	15	15
7	10	9	13	7	12	9	13	7	0	0
8	44	38	48	40	40	43	37	31	12	7
9	29	25	31	27	28	32	19	16	10	0
10	12	21	18	11	25	17	14	6	0	0
11	5	2	3	2	2	0	0	0	0	0
12	44	36	43	38	26	22	15	12	0	0
13	3	2	0	0	0	0	0	0	0	0
14	3	11	11	0	17	10	0	0	0	0
15	22	18	20	9	10	7	5	0	0	0
16	45	36	28	26	30	13	1	0	0	0
17	6	1	0	0	0	0	0	0	0	0
18	29	29	16	27	6	6	0	0	0	0
19	2	0	0	0	0	0	0	0	0	0
20	30	16	13	19	12	3	0	0	0	0
21	11	0	10	10	0	1	0	0	0	0
22	1	6	0	0	3	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	21	32	28	9	14	3	0	0	0	0
25	3	0	0	0	0	0	0	0	0	0
26	0	1	0	0	0	0	0	0	0	0
27	5	0	3	0	0	0	0	0	0	0
28	14	11	11	0	3	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
30	17	12	16	7	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0
32	28	16	19	6	0	0	0	0	0	0

Table G.1 – Steady-state harmonic analysis data for Doncaster Parish Church

G.2 Analysis data for Heslington Parish Church

Harmonic	C	F#	c	f#	c1	f#1	c2	f#2	c3	f#3
1	35	45	47	49	53	38	43	45	38	45
2	29	50	35	53	47	45	46	49	46	41
3	28	30	33	35	33	40	33	35	33	31
4	40	39	43	50	38	45	41	42	36	43
5	15	14	14	23	14	12	20	20	15	4
6	37	39	41	29	32	37	32	28	27	9
7	13	4	14	13	12	11	10	2	4	0
8	34	34	34	35	39	37	25	27	11	13
9	6	5	14	8	13	24	15	5	9	0
10	21	28	22	23	12	15	18	2	3	0
11	0	5	2	2	0	0	0	0	0	0
12	38	35	42	39	27	30	24	13	7	0
13	10	9	4	0	0	0	0	0	0	0
14	0	10	10	3	9	2	0	0	0	0
15	3	7	0	0	0	0	0	0	0	0
16	31	20	37	17	27	12	10	6	0	0
17	7	0	0	0	0	0	0	0	0	0
18	22	19	23	19	5	0	2	0	0	0
19	3	0	0	0	0	0	0	0	0	0
20	16	10	7	6	11	2	3	0	0	0
21	0	0	0	0	0	0	0	0	0	0
22	0	2	0	1	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	31	33	26	11	13	5	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0
26	1	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0
28	10	2	1	1	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
30	9	3	7	1	0	1	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0
32	20	15	17	11	0	0	0	0	0	0

Table G.2 – Steady-state harmonic analysis data for Heslington Parish Church

G.3 Analysis data for Jack Lyons

Harmonic	C	F#	c	f#	c1	f#1	c2	f#2	c3	f#3
1	27	28	43	48	47	38	45	38	36	39
2	50	55	42	44	37	43	52	45	47	42
3	27	40	40	23	38	40	33	37	38	31
4	42	50	51	41	45	48	48	44	45	46
5	16	28	11	21	8	11	22	0	15	0
6	33	42	41	38	35	45	40	42	16	23
7	11	21	24	14	7	2	9	2	0	11
8	50	43	34	26	45	33	39	39	17	27
9	13	9	12	12	6	17	11	8	8	4
10	12	10	6	12	16	7	12	5	9	0
11	8	7	5	2	4	2	0	0	0	0
12	43	43	50	45	43	40	30	12	1	0
13	9	4	5	0	0	0	0	0	0	0
14	4	10	14	11	22	8	0	8	0	0
15	6	0	1	0	1	5	0	3	0	0
16	43	42	46	42	27	5	11	6	0	0
17	4	5	1	0	0	0	0	0	0	0
18	26	20	13	21	18	13	4	0	0	0
19	0	0	4	0	0	0	0	0	0	0
20	12	13	4	7	1	6	3	0	0	0
21	0	0	0	0	0	0	0	0	0	0
22	6	0	0	10	5	4	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	31	42	28	16	13	18	0	0	0	0
25	0	1	0	0	0	0	0	0	0	0
26	7	1	9	1	0	0	0	0	0	0
27	0	0	3	0	0	0	0	0	0	0
28	4	13	8	6	7	4	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
30	4	9	4	0	0	6	0	0	0	0
31	1	0	0	0	0	0	0	0	0	0
32	42	30	23	21	4	13	0	0	0	0

Table G.3 – Steady-state harmonic analysis data for the Jack Lyons concert hall

G.4 Analysis data for Port Sunlight URC

Harmonic	C	F#	c	f#	c1	f#1	c2	f#2	c3	f#3
1	36	41	45	52	51	36	45	38	41	37
2	41	42	36	42	34	42	38	44	39	43
3	41	43	27	27	33	34	23	29	31	19
4	42	32	31	39	31	36	34	32	34	24
5	33	29	22	21	17	14	16	15	8	0
6	29	28	13	31	30	31	19	20	18	0
7	18	27	12	21	10	9	12	0	5	0
8	35	37	27	28	33	33	25	15	10	3
9	16	7	11	18	6	16	10	4	0	0
10	17	12	6	12	13	5	0	0	2	0
11	12	0	5	0	0	0	0	0	0	0
12	28	17	22	26	20	13	20	1	0	0
13	6	0	0	0	0	0	0	0	0	0
14	10	10	4	5	0	2	0	0	0	0
15	3	0	0	0	0	0	0	0	0	0
16	26	26	26	24	10	6	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0
18	0	11	13	17	3	4	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0
20	6	10	2	10	0	0	2	0	0	0
21	1	0	0	0	0	0	0	0	0	0
22	0	1	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	1	11	12	11	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0
28	0	5	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0
30	0	0	1	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0
32	7	13	12	0	0	0	0	0	0	0

Table G.4 – Steady-state harmonic analysis data for Port Sunlight URC

G.5 Analysis data for St Chad's (without mixture)

Harmonic	C	F#	c	f#	c1	f#1	c2	f#2	c3	f#3
1	48	38	49	52	38	42	47	41	45	43
2	51	42	51	51	50	52	47	51	44	42
3	42	42	30	35	41	38	32	33	23	40
4	46	54	48	49	45	42	46	39	43	37
5	12	25	29	31	25	23	29	28	16	5
6	48	49	47	41	36	38	13	33	17	13
7	18	29	26	18	2	16	16	8	5	4
8	42	37	42	44	31	25	36	30	13	8
9	14	26	19	10	12	15	3	0	0	0
10	37	34	30	23	22	9	10	7	12	0
11	10	15	7	8	4	9	0	0	0	0
12	36	25	34	24	29	27	14	16	5	0
13	1	5	5	0	2	0	0	0	0	0
14	8	24	18	13	4	0	0	8	0	0
15	5	7	0	0	0	0	0	0	0	0
16	23	18	35	27	23	12	6	3	0	0
17	0	0	0	0	4	0	0	0	0	0
18	6	13	1	0	8	4	0	0	0	0
19	3	2	0	0	0	0	0	0	0	0
20	25	16	6	18	4	12	3	0	0	0
21	10	0	0	0	0	0	0	0	0	0
22	11	10	2	1,4	2	2	0	0	0	0
23	3	1	0	0	0	0	0	0	0	0
24	7	16	7	10	7	6	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0
26	3	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0
28	17	9	18	2	0	0	0	0	0	0
29	4	1	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0
32	4	5	5	0	0	0	0	0	0	0

Table G.5 – Steady-state harmonic analysis data for St Chad's (without mixture)

G.6 Analysis data for St Chad's (with mixture)

Harmonic	C	F#	c	f#	c1	f#1	c2	f#2	c3	f#3
1	42	38	49	52	35	43	24	47	46	42
2	55	38	49	51	51	53	32	54	45	46
3	37	42	38	37	35	47	41	40	36	34
4	48	54	43	55	41	51	47	46	45	40
5	42	42	37	47	44	40	31	29	12	1
6	45	50	40	43	32	28	35	38	14	16
7	23	29	20	19	9	12	12	11	1	3
8	45	43	19	45	41	21	34	36	14	20
9	10	25	16	10	29	24	14	21	0	2
10	28	29	39	18	23	22	6	4	13	0
11	11	14	5	2	5	8	1	0	0	0
12	34	41	41	31	27	32	28	17	12	0
13	9	7	4	0	2	0	0	0	0	0
14	17	17	20	12	4	10	4	10	0	0
15	29	23	28	19	25	15	0	0	0	0
16	26	22	37	14	25	3	12	9	0	0
17	3	4	0	0	0	0	0	0	0	0
18	32	22	28	15	11	7	0	0	0	0
19	8	4	0	0	0	0	0	0	0	0
20	26	22	6	7	17	8	9	0	0	0
21	5	2	8	0	0	0	0	0	0	0
22	10	4	2	4	5	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0
24	23	15	21	7	10	1	0	0	0	0
25	20	13	11	6	7	1	0	0	0	0
26	6	3	0	0	0	0	0	0	0	0
27	0	1	0	0	0	0	0	0	0	0
28	11	15	19	1	0	0	0	0	0	0
29	3	0	1	0	0	0	0	0	0	0
30	21	16	5	0	2	0	0	0	0	0
31	6	0	0	0	0	0	0	0	0	0
32	11	17	11	6	5	0	0	0	0	0

Table G.6 – Steady-state harmonic analysis data for St Chad's (with mixture)

Appendix H: Specialised listener results

This appendix contains the specialised listener results referred to in section 5.6.

Identity	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Age	y	y	y	o	y	o	o	o	o	y	y	y	y	y	y	o	o	o	y	y	y	o
Nation	UK	UK	US	Can	US	US	US	US	US	UK	UK	UK	UK	US	US	UK	UK	UK	US	US	US	Can
Test 1																						
thin	7	10	8	9	7	9	10	6	3	6	6	8	1	4	8	4	4	2	9	7	8	8
flutey	7	2	2	3	5	2	4	6	6	4	3	8	3	4	4	4	7	2	2	6	6	5
full	7	2	10	3	9	2	2	10	9	9	10	2	8	8	4	8	9	10	9	8	3	4
warm	3	2	2	3	4	2	3	3	4	3	2	2	5	4	3	8	6	2	2	4	3	4
balanced	4	2	2	9	3	9	4	7	9	5	3	5	3	5	4	8	8	6	2	7	3	5
Test 2																						
thin	6	5	8	3	4	1	6	7	5	6	5	5	6	3	4	8	4	10	9	5	7	7
flutey	7	8	3	10	8	6	10	8	6	4	6	5	6	8	7	5	8	10	10	9	7	5
full	8	8	9	3	7	10	7	3	9	4	4	7	9	9	5	8	9	2	3	4	5	4
warm	5	9	9	8	9	10	7	7	11	9	7	8	2	9	9	4	8	2	9	4	5	4
balanced	3	8	10	3	10	10	3	7	11	4	5	6	4	9	4	7	8	2	8	7	7	6
Test 3																						
thin	5	4	8	9	4	6	5	4	6	4	7	3	4	6	4	5	8	6	9	8	5	5
flutey	5	6	8	3	6	6	9	3	6	7	7	3	6	9	6	9	4	2	9	6	7	5
full	8	4	3	8	8	6	8	7	8	4	3	7	4	3	4	8	4	10	9	3	7	7
warm	5	8	8	9	9	9	8	5	7	8	8	8	9	9	3	8	7	10	9	6	7	7
balanced	6	5	4	4	3	10	4	3	4	3	3	8	9	9	8	7	5	10	7	4	5	8
Test 4																						
thin	5	2	9	2	2	2	1	6	5	7	6	2	3	6	3	5	5	6	2	5	2	6
flutey	7	9	10	9	10	9	11	9	8	8	8	3	8	9	9	9	4	10	10	8	7	6
full	6	9	10	3	3	9	8	4	3	2	1	8	3	4	4	6	5	2	9	4	7	3
warm	8	9	10	9	9	9	9	7	8	10	8	8	9	9	9	9	8	2	10	8	7	4
balanced	5	9	10	4	8	9	6	6	4	4	8	7	10	8	7	7	5	6	8	4	7	4
Test 5																						
thin	9	10	3	9	9	10	11	6	9	5	5	8	6	6	9	8	8	6	9	7	9	6
flutey	9	6	2	6	4	6	1	7	5	3	5	8	6	4	5	4	7	2	3	5	5	7
full	5	7	9	8	8	2	2	7	8	10	10	4	8	9	8	7	7	2	2	8	5	7
warm	3	5	4	3	3	2	2	3	4	4	4	4	3	3	5	3	5	2	2	4	3	4
balanced	4	5	9	8	7	10	4	7	8	7	3	3	4	4	4	4	7	6	4	8	4	6
Test 6																						
thin	5	7	5	10	8	10	9	4	8	4	7	9	8	4	8	8	8	2	7	7	5	6
flutey	5	6	4	2	5	2	1	2	5	3	6	9	2	4	3	3	5	2	2	5	6	7
full	8	5	3	8	9	9	8	8	2	9	8	4	9	9	5	7	5	10	9	7	5	5
warm	6	5	2	3	10	3	3	4	1	4	3	4	2	4	4	3	5	10	2	5	5	5
balanced	8	7	1	9	11	9	9	6	4	8	5	8	3	6	8	7	7	6	8	7	5	5

Table H.1 – Specialised listener results

Appendix I: Matlab scripts

This appendix includes all Matlab scripts referred to in the text. Lines of code beginning with a percentage symbol are comments.

I.1 Note analysis script

This script is one of ten used to analyse each note in section 5.7.1. Other scripts were identical apart from the pitch-related components.

```
% "c1" generates a harmonic spectral analysis of an input signal of approx. pitch 256Hz

% Filename to be analysed.
filename = 'c1.wav'

% Upper frequency cutoff point - the maximum frequency in Hz which will be displayed on
% the screen.  The range depends on what you're studying - a sensible setting for a
% 250Hz signal with up to 20 harmonics would be 5000-6000Hz.  You can't set the cutoff
% to be more than the Nyquist frequency, which is half the sampling rate - so for a CD
% quality recording sampled at 44.1kHz, the maximum cutoff would be 22050Hz.
ufcop=8500;

% dB cutoff point.  This adjusts the minimum amplitude at which harmonics will be
% displayed. To see quieter harmonics, reduce the level.  If there is a lot of noise,
% try increasing the level.  Sensible range is from -80 (sensitive) to 0 (insensitive).
dbcutoff=-50;

% FFT Windowsize - 4096 seems quite reasonable.  The larger the window size, the better
% the FFT is at identifying frequency components, but the more calculation it must do.
% The window size can't be more than the number of samples, so a tenth of a second file
% can have a maximum window size of 4410.  If the windowsize is accidentally bigger
% than the number of samples, it will be repeatedly halved until it isn't.
windowsize = 4096;

% Window type to apply: 0 = no window, 1 = hamming, 2 = hanning.  A window gives some
% samples more emphasis than others.  Choose one and stick to it for a particular
% project, unless you have very good reason to change, as results from one method
% cannot be precisely compared with those from the other or no windowing function.
windowtype = 1;

% Method of finding harmonics.  Both try to identify the fundamental.  Method 1 picks
% out the first point above the dB offset point. Method 2 averages the waveform,
% picking out the first that is more than the dB offset above the average (more useful
% for low frequency noise?). The minimum frequency should be set to just below the
% frequency of the likely lowest note.  This is again to prevent low frequency noise
% interfering with the low note identification. Subsequent harmonics are identified
% whether they're above or below the dB offset.  More methods may be added if these
% prove unreliable.  Harmonics to find indicates the maximum number of harmonics to
% find (must be at least 2).  The error margin is the amount below the expected
```

```
% harmonic that searching for it is begun at.  Adjust only if prominent harmonics are
% clearly being missed - sensible range 0.1 to 0.5
harmonicmethod=1;
dboffset=40;
minfreq=250;
harmonicstofind=32;
errormargin=0.25;

% --- No user alterable functions beyond this point ---

% dBcutoff on/off button - must be on for harmonics detection to work.
usedbcutoff=1;

% Automatically adjust amplitude dB scale to dB cutoff point - must be on for
% harmonics.
autoadjust=1;

% Real file read in.  Assume mono for now.  Auto adjusts to sampling frequency

[y,Fs] = WAVREA(filename);

% Check that fft window isn't greater than the sample size.

while size(y,1) < windowsize
windowsize = (windowsize/2);
end
str='FFT window size is ';
disp(str);
disp(windowsize);

% Apply a windowing function

if windowtype > 0
      if windowtype == 1
                              windowfunc = hamming(windowsize);
else windowfunc = hanning(windowsize);
                     end

for counterone = 1:windowsize
                              y(counterone) = (windowfunc(counterone)*y(counterone));
                     end
end

% Do the fft of the thing

Y = fft(y,windowsize);

% Get power spectral density

Pyy = 10*log10(Y.* conj(Y) / windowsize);
```

```
% Check that frequency cutoff isn't too big:

if ufcop > (Fs/2)
         ufcop = (Fs/2);
end

% Calculate proportion of fft to display (also halve it 'cos of symmetry)

displaywindow = windowsize*(ufcop/Fs);

% Apply dB cutoff point.  This is after the display window to avoid unnecessary
calculation.
% Also, if enabled, auto adjust to reference from 0dB

if usedbcutoff == 1
    for counterone = 1:displaywindow
        if Pyy(counterone) < dbcutoff
            Pyy(counterone) = dbcutoff;
        end
        if autoadjust == 1
            Pyy(counterone) = Pyy(counterone) - dbcutoff;
        end
    end
end

% Set up matrix of the harmonics

Hyy=zeros(size(Pyy));

% Find first harmonic

startpoint=floor(minfreq/(Fs/windowsize));

if harmonicmethod == 2
    for counterone = startpoint:displaywindow
        avglevel = avglevel + (Pyy(counterone)/displaywindow);
    end
    else avglevel=0;
end

harmonicfound = 0;
freqnumber = startpoint;

while harmonicfound == 0
    freqnumber = (freqnumber+1);
    if freqnumber > size(Pyy)
        harmonicfound=1;
        freqnumber=(minfreq/(Fs/windowsize));
    end
    if Pyy(freqnumber) > avglevel + dboffset + dbcutoff
        if Pyy(freqnumber+1) < Pyy(freqnumber)
```

```
            harmonicfound = 1;
        end
    end
end

% Find and enter harmonics in the matrix.  This function looks for the next harmonic,
% basing its search upon the position of the previous one.  It allows for a given
% margin of error (set elsewhere) from the likely position - this is because at low
% resolutions, the first harmonic is approximated to within a large margin of error.
% Once it has found the harmonic, its amplitude is entered in the matrix.  If a
% harmonic is not found, the value -10 is entered into the matrix.

Hyy(freqnumber) = Pyy(freqnumber);
disp(freqnumber);
for counterone = 2:harmonicstofind
    startposition = floor(freqnumber *(1+ ((1-errormargin) * (1/(counterone-1))) ));
    endposition = ceil(freqnumber *(1+ ((1+errormargin) * (1/(counterone-1))) ));
    [maxpt, maxptloc] = max(Pyy(startposition:endposition));
    maxptloc = (maxptloc + startposition-1);
        if maxpt > 0
            Hyy(maxptloc) = Pyy(maxptloc);
        else
            maxptloc = round((startposition+endposition)/2);
            Hyy(maxptloc) = -10;
        end
    disp(Hyy(maxptloc));
    freqnumber=maxptloc;
end

% Plot it

f = Fs*(0:(displaywindow-1))/windowsize;
plot(f,Pyy(1:displaywindow),'y-',f,Hyy(1:displaywindow),'c-');
```

I.2 Chordal analysis script

This script analyses the chords synthesised in section 6.1.2 for spectral centroid, smoothness and slope.

```
%Script to generate fft, spectral centroid and spectral slope data for analysis.

% Filename to be analysed.
filename = 's1.wav'

% Upper frequency cutoff point - the maximum frequency in Hz which will be displayed on
% the screen.  The range depends on what you're studying - a sensible setting for a
% 250Hz signal with up to 20 harmonics would be 5000-6000Hz.  You can't set the cutoff
% to be more than the Nyquist frequency, which is half the sampling rate - so for a CD
% quality recording sampled at 44.1kHz, the maximum cutoff would be 22050Hz.
ufcop=11025;

% FFT Windowsize - 4096 seems quite reasonable.  The larger the window size, the better
% the FFT is at identifying frequency components, but the more calculation it must do.
% The window size can't be more than the number of samples, so a tenth of a second file
% can have a maximum window size of 4410.  If the windowsize is accidentally bigger
% than the number of samples, it will be repeatedly halved until it isn't.
windowsize = 8192;

% Window type to apply: 0 = no window, 1 = hamming, 2 = hanning.  A window gives some
% samples more emphasis than others.  Choose one and stick to it for a particular
% project, unless you have very good reason to change, as results from one method
% cannot be precisely compared with those from the other or no windowing function.
windowtype = 1;

% dB cutoff point.  This adjusts the minimum amplitude at which harmonics will be
% displayed. To see quieter harmonics, reduce the level.  If there is a lot of noise,
% try increasing the level.  Sensible range is from -80 (sensitive) to 0 (insensitive).
dbcutoff=-80;


% dBcutoff on/off button.
usedbcutoff=1;
% Automatically adjust amplitude dB scale to dB cutoff point.
autoadjust=1;

% Real file read in.  Assume mono for now.  Auto adjusts to sampling frequency

[y,Fs] = WAVREA(filename);

% Check that fft window isn't greater than the sample size.

while size(y,1) < windowsize
windowsize = (windowsize/2);
end
str='FFT window size is ';
```

```
disp(str);
disp(windowsize);

% Apply a windowing function

if windowtype > 0
    if windowtype == 1
        windowfunc = hamming(windowsize);
    else windowfunc = hanning(windowsize);
    end

    for counterone = 1:windowsize
        y(counterone) = (windowfunc(counterone)*y(counterone));
    end
end

% Do the fft of the thing

Y = fft(y,windowsize);

% Get power spectral density

Pyy = 10*log10(Y.* conj(Y) / windowsize);

% Check that frequency cutoff isn't too big:

if ufcop > (Fs/2)
    ufcop = (Fs/2);
end

% Calculate proportion of fft to display (also halve it 'cos of symmetry)

displaywindow = windowsize*(ufcop/Fs);

% Apply dB cutoff point.  This is after the display window to avoid unnecessary
% calculation. Also, if enabled, auto adjust to reference from 0dB

if usedbcutoff == 1
    for counterone = 1:displaywindow
        if Pyy(counterone) < dbcutoff
            Pyy(counterone) = dbcutoff;
        end
        if autoadjust == 1
            Pyy(counterone) = Pyy(counterone) - dbcutoff;
        end
    end
end

% Get spectral centroid

topval=0;
```

```
botval=0;
for counterone = 1:displaywindow
    topval=(topval+(counterone*Pyy(counterone)));
    botval=(botval+Pyy(counterone));
end
centval=topval/botval;
centval=(centval*(ufcop/displaywindow));
str='Spectral centroid is ';
disp(str);
disp(centval);

% Get spectral slope
specslop=0;
for counterone = 1:(displaywindow/2)
    specslop=(specslop+Pyy(counterone));
end
for counterone = (displaywindow/2)+1:displaywindow
    specslop=(specslop-Pyy(counterone));
end
specslop=(2*specslop/displaywindow);
str='Spectral slope is ';
disp(str);
disp(specslop);

% Get spectral smoothness
specsmooth=0;
for counterone = 2:displaywindow
    specsmooth=(specsmooth+abs(Pyy(counterone)-Pyy(counterone-1)));
end
specsmooth=(specsmooth/displaywindow);
str='Spectral smoothness is ';
disp(str);
disp(specsmooth);

% Plot it

f = Fs*(0:(displaywindow-1))/windowsize;
plot(f,Pyy(1:displaywindow));
```

Appendix J: Comparative acoustic measures

This appendix includes data for the graphs comparing organ ensembles in section 5.7.2 derived from the data in appendix G. In all these tables, "average" is a simple numerical avererage of the lowest eight data points, "slope" is the average of the highest four subtracted from the average of the lowest four, and "smooth" is the average difference between one data point and the next. "stdev" is the standard deviation, where appropriate to include this measure, and has no units. Unless otherwise stated, all other measures are in decibels.

J.1 Normalised spectral centroid

Pitch	Port Sunlight	Jack Lyons	Heslington	Doncaster	St Chad (mix)	St Chad (no mix)
C	7.363	11.985	11.416	12.211	11.598	9.416
F#	8.366	11.196	9.784	11.705	10.804	8.978
c	8.367	10.538	9.915	10.904	10.474	8.301
f#	7.430	10.004	7.888	9.056	7.681	7.282
c1	5.426	8.608	7.265	8.286	8.699	6.987
f#1	4.153	6.730	4.819	5.201	5.015	5.023
c2	4.811	5.716	5.649	5.316	6.306	4.950
f#2	3.556	5.232	4.493	4.710	5.287	4.896
c3	3.356	3.901	4.022	3.707	4.038	3.924
f#3	2.381	3.888	3.065	3.058	3.412	3.073
Average	6.184	8.751	7.654	8.424	8.233	6.979
Slope	3.395	4.359	4.194	5.091	3.813	3.030
Smooth	1.019	0.965	1.264	1.104	1.561	0.646

Table J.1 – Normalised spectral centroid data (in normalised frequency units)

J.2 Average harmonic strength

Pitch	Port Sunlight	Jack Lyons	Heslington	Doncaster	St Chad (mix)	St Chad (no mix)
C	12.75	16.59	14.34	18.63	21.44	16.69
F#	12.56	17.69	14.31	15.72	20.38	16.97
c	10.22	16.31	14.78	17.31	18.66	15.91
f#	12.00	14.41	13.41	14.50	15.65	14.29
c1	9.09	13.41	12.03	13.66	15.16	12.16
f#1	8.78	12.75	11.13	12.06	13.31	11.63
c2	7.63	11.22	10.06	10.06	10.31	9.44
f#2	8.25	12.04	11.42	11.79	15.08	12.38
c3	11.75	14.5	14.31	12.38	14.88	13.94
f#3	10.50	18.58	15.50	14.42	17.00	16.00
Average	10.160	14.303	12.685	14.216	16.249	13.684
Slope	3.445	3.895	3.050	4.648	5.568	4.563
Smooth	1.329	1.199	0.940	1.926	2.271	1.536

Table J.2 – Average harmonic strength data

J.3 Spectral smoothness

Pitch	Port Sunlight	Jack Lyons	Heslington	Doncaster	St Chad (mix)	St Chad (no mix)
C	6.16	18.55	15.84	18.68	15.13	14.19
F#	8.71	18.00	15.16	16.65	12.61	11.51
c	7.90	15.55	15.74	16.19	13.55	12.96
f#	9.55	15.39	12.13	13.00	10.52	11.38
c1	6.55	15.45	10.87	11.55	10.77	9.81
f#1	5.94	12.94	8.65	9.16	6.61	7.29
c2	4.61	9.13	6.55	5.00	7.29	5.84
f#2	4.26	11.74	7.43	6.61	9.26	8.48
c3	5.40	7.47	7.47	5.80	9.60	9.13
f#3	5.91	13.91	9.54	5.73	11.45	6.09
Average	6.710	14.594	11.546	12.105	10.718	10.183
Slope	2.740	4.558	6.343	8.050	4.470	4.655
Stdev	1.883	3.168	3.783	4.943	2.973	2.859
Smooth	1.471	1.736	1.619	2.184	1.936	1.984

Table J.3 – Spectral smoothness data

J.4 Spectral slope

Pitch	Port Sunlight	Jack Lyons	Heslington	Doncaster	St Chad (mix)	St Chad (no mix)
C	23.63	16.06	13.81	16.38	19.75	21.75
F#	18.75	18.63	18.13	15.94	23.50	24.81
c	15.44	20.50	19.44	20.13	23.31	26.94
f#	19.25	18.56	20.56	19.25	25.56	24.66
c1	17.81	20.81	20.44	22.56	23.19	21.19
f#1	17.06	17.50	21.25	22.50	24.50	20.25
c2	15.00	21.56	19.50	20.13	19.50	18.50
f#2	16.50	21.25	21.83	23.58	27.00	22.92
c3	23.00	24.50	23.88	22.25	23.50	23.63
f#3	20.00	23.17	26.67	26.50	25.67	28.00
Average	17.930	19.359	19.370	20.059	23.289	22.628
Stdev	2.739	1.976	2.525	2.823	2.600	2.754

Table J.4 – Average harmonic strength data

J.5 Inter-quartile spectral slope

Pitch	Port Sunlight	Jack Lyons	Heslington	Doncaster	St Chad (mix)	St Chad (no mix)
C	13.75	7.38	3.75	7.88	7.13	8.63
F#	4.88	5.63	6.88	8.38	13.13	12.00
c	5.88	11.13	9.38	10.88	15.63	16.5
f#	5.88	8.75	6.88	6.00	9.13	9.45
c1	5.75	10.25	7.38	12.88	12.13	8.88
f#1	4.75	5.38	9.50	11.00	12.25	6.00
c2	3.50	7.13	7.75	6.75	7.00	3.75
f#2	3.33	8.17	7.17	12.00	11.67	8.33
c3	9.75	7.50	9.50	9.25	4.00	8.50
f#3	7.00	9.00	14.33	15.33	10.67	14.33
Average	5.965	7.978	7.336	9.471	11.009	9.193
Stdev	3.302	2.038	1.786	2.550	3.017	3.823

Table J.5 – Average harmonic strength data

Appendix K: Acoustic data of synthesised ensembles

This appendix presents the acoustic data of harmonic strength levels for the ensembles synthesised in chapter six. Sections K.1 to K.5 present the individual data for each stop used at each of six voicing points. Sections K.6 to K.8 present the composite data for each of the ensembles analysed in section 6.1.2. In all cases, where a harmonic is omitted in the left column, all values across the table were zero. Harmonic numbers are always with respect to those of the 8' unison.

K.1 Harmonic amplitude data for the Open Diapason stop

Harmonic	C	c	c1	c2	c3	c4
1	225	221	214	209	203	191
2	200	199	190	197	185	154
3	191	173	152	156	150	120
4	145	154	137	138	130	82
5	129	139	117	115	115	52
6	130	133	107	115	91	30
7	132	117	101	104	59	
8	105	99	84	100	54	
9	105	92	61	81	40	
10	93	88	69	72		
11	90	76	60	69		
12	80	70	49	47		
13	78	56	27	30		
14	57	43		24		
15	72	34				
16	65					
17	67					
18	23					
19	46					
20	52					

Table K.1 – Harmonic amplitude of the Open Diapason

K.2 Harmonic amplitude data for the Stopped Diapason stop

Harmonic	C	c	c1	c2	c3	c4
1	204	192	197	198	183	169
2	81	112	99	102	47	47
3	160	152	129	132	89	78
4	68	51	45	48	56	0
5	98	99	88	91	88	5
6	63	18	49	52	42	
7	100	63	69	72	58	
8	38	16	37	40	6	
9	58	56	48	51	24	
10	18	0	6	9	12	
11	25	12	4	7		
12	13					
13	8					

Table K.2 – Harmonic amplitude of the Stopped Diapason

K.3 Harmonic amplitude data for the Principal stop

Harmonic	C	c	c1	c2	c3	c4
2	204	198	198	185	178	176
4	193	187	187	174	167	165
6	151	145	145	132	125	123
8	134	128	128	115	108	
10	105	99	99	86	79	
12	114	108	108	95	88	
14	67	61	61	48		
16	90	84	84	71		
18	41	35	35	22		
20	69	63	63	50		

Table K.3 – Harmonic amplitude of the Open Diapason

K.4 Harmonic amplitude data for the Fifteenth stop

Harmonic	C	c	c1	c2	c3	c4
4	207	208	201	188	196	198
8	188	189	180	156	164	97
12	162	159	158	136	125	
16	142	141	128	122		
20	113	137	105			
24	101	127	87			
28	72	98	79			
32	86	97	53			
36	76	75	29			
40	91	81	42			
44	73	68				
48	78	78				
52	53	60				
56	64	68				
60	56	55				
64	66	54				
68	55	30				
72	51	29				
76	47	28				
80	42	38				

Table K.4 – Harmonic amplitude of the Fifteenth

K.5 Harmonic amplitude data for the Mixture stop

Harmonic	C	c	c1	c2	c3	c4
2	0	0	0	0	36	201
3	0	0	28	180	189	193
4	0	17	185	184	193	375
6	176	178	177	361	347	513
8	180	182	374	347	355	203
9	0	0	0	132	145	145
12	351	334	486	408	429	316
16	358	351	277	280	288	
18	124	146	129	142	142	
20	0	0	100	0		
24	419	413	511	256		
28	0	0	74	0		
30	138	93	0	0		
32	289	275	216	114		
36	267	231	139			
40	142	97	96			
42	73	67				
48	468	454				
54	106	0				
56	77	71				
60	172	89				
64	217	205				
72	212	0				
80	86	93				
84	64					
96	87					
120	86					

Table K.5 – Harmonic amplitude of the Mixture

K.6 Harmonic amplitude data for the 842 ensemble

Harmonic	C	c	c1	c2	c3	c4
1	225	221	214	209	203	191
2	404	397	388	382	363	330
3	191	173	152	156	150	120
4	545	549	525	500	493	445
5	129	139	117	115	115	52
6	281	278	252	247	216	153
7	132	117	101	104	59	0
8	427	416	392	371	326	97
9	105	92	61	81	40	
10	198	187	168	158	79	
11	90	76	60	69	0	
12	356	337	315	278	213	
13	78	56	27	30		
14	124	104	61	72		
15	72	34	0	0		
16	297	225	212	193		
17	67	0	0	0		
18	64	35	35	22		
19	46	0	0	0		
20	234	200	168	50		
24	101	127	87			
28	72	98	79			
32	86	97	53			
36	76	75	29			
40	91	81	42			
44	73	68				
48	78	78				
52	53	60				
56	64	68				
60	56	55				
64	66	54				
68	55	30				
72	51	29				
76	47	28				
80	42	38				

Table K.6 – Harmonic amplitude of the 842 ensemble

K.7 Harmonic amplitude data for the S842 ensemble

Harmonic	C	c	c1	c2	c3	c4
1	204	192	197	198	183	169
2	285	310	297	287	225	223
3	160	152	129	132	89	78
4	468	446	433	410	419	363
5	98	99	88	91	88	5
6	214	163	194	184	167	123
7	100	63	69	72	58	0
8	360	333	345	311	278	97
9	58	56	48	51	24	
10	123	99	105	95	91	
11	25	12	4	7	0	
12	289	267	266	231	213	
13	8	0	0	0		
14	67	61	61	48		
16	232	225	212	193		
18	41	35	35	22		
20	182	200	168	50		
24	101	127	87			
28	72	98	79			
32	86	97	53			
36	76	75	29			
40	91	81	42			
44	73	68				
48	78	78				
52	53	60				
56	64	68				
60	56	55				
64	66	54				
68	55	30				
72	51	29				
76	47	28				
80	42	38				

Table K.7 – Harmonic amplitude of the S842 ensemble

K.8 Harmonic amplitude data for the 842M ensemble

Harmonic	C	c	c1	c2	c3	c4
1	225	221	214	209	203	191
2	404	397	388	382	399	531
3	191	173	180	336	339	313
4	545	566	710	684	686	820
5	129	139	117	115	115	52
6	457	456	429	608	563	666
7	132	117	101	104	59	0
8	607	598	766	718	681	300
9	105	92	61	213	185	145
10	198	187	168	158	79	0
11	90	76	60	69	0	0
12	707	671	801	686	642	316
13	78	56	27	30	0	
14	124	104	61	72	0	
15	72	34	0	0	0	
16	655	576	489	473	288	
17	67	0	0	0	0	
18	188	181	164	164	142	
19	46	0	0	0		
20	234	200	268	50		
24	520	540	598	256		
28	72	98	153	0		
30	138	93	0	0		
32	375	372	269	114		
36	343	306	168			
40	233	178	138			
42	73	67				
44	73	68				
48	546	532				
52	53	60				
54	106	0				
56	141	139				
60	228	144				
64	283	259				
68	55	30				
72	263	29				
76	47	28				
80	128	131				
84	64					
96	87					
120	86					

Table K.8 – Harmonic amplitude of the 842M ensemble

Appendix L: Synthesised listening test results

This appendix contains the synthesised listener test results referred to in section 6.2.

Identity	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Age	u	o	o	u	o	u	u	o	o	u	u	o	?	o	o	u
Nation	UK	UK	US	UK	UK	UK	UK	UK	Can	US	US	US	UK	US	US	US
Test 1																
thin	8	3	4	8	7	4	9	6	5	7	8	3	8	7	2	4
flutey	4	5	6	3	4	1	5	9	6	4	3	2	3	4	3	3
bright	10	9	10	11	8	8	9	11	8	11	10	10	9	8	6	9
warm	3	4	7	1	4	7	3	4	4	3	4	6	3	4	9	4
clear	8	9	9	7	8	8	9	10	8	8	4	11	6	7	6	3
Test 2																
thin	4	4	3	2	4	5	4	3	8	3	4	6	1	4	6	3
flutey	5	10	7	7	8	6	4	6	7	8	9	5	8	8	6	9
bright	4	5	4	3	6	4	5	8	5	3	5	7	6	5	5	4
warm	8	9	9	9	8	6	7	5	8	7	10	6	10	7	8	7
clear	3	5	7	3	5	5	5	3	5	7	8	8	4	6	5	4
Test 3																
thin	7	8	8	7	8	6	2	6	7	8	9	2	8	5	9	4
flutey	8	8	8	7	8	7	8	11	8	6	8	2	8	7	8	8
bright	3	3	3	2	4	3	3	1	5	8	3	11	3	5	6	3
warm	6	8	8	7	3	7	8	11	5	4	9	8	9	8	8	7
clear	5	2	6	3	4	4	8	1	7	6	5	10	6	6	6	9
Test 4																
thin	7	8	6	7	4	7	7	6	6	4	3	3	7	7	7	6
flutey	5	5	8	6	6	6	7	10	7	7	2	3	8	5	6	6
bright	6	8	4	7	5	4	7	3	8	7	8	10	3	7	6	6
warm	6	7	8	2	8	6	8	9	5	5	5	7	4	4	4	5
clear	8	7	4	8	9	5	7	10	5	6	4	8	8	7	6	7
Test 5																
thin	4	5	8	5	8	5	3	6	5	6	8	9	6	7	8	4
flutey	8	7	8	9	8	10	7	11	7	7	9	9	9	8	7	8
bright	2	2	3	2	3	1	3	1	4	4	2	2	2	4	2	3
warm	9	8	7	9	8	8	8	10	7	8	8	8	8	7	7	9
clear	4	3	3	3	6	4	4	1	7	5	5	2	3	6	6	8
Test 6																
thin	7	9	8	8	9	9	7	7	8	8	7	9	9	7	7	7
flutey	7	4	5	4	8	8	6	8	8	5	4	6	7	6	6	3
bright	7	4	6	6	8	4	7	4	4	7	8	3	4	5	5	9
warm	5	3	4	7	5	8	4	7	6	7	4	6	2	6	6	4
clear	8	6	5	7	7	4	5	5	4	6	7	3	3	5	6	8

Table L.1 – Synthesised listening test results

Glossary

BMIS:	The Bradford Musical Instrument Synthesiser.
DEV:	Digital Enhanced Voicing, the Bradford voicing software.
DPAT:	Duration of Perceived Attack Time.
Flue:	A common kind of organ pipe, which works much like a recorder.
Foundation:	Stops of unison pitch to which other stops of non-unison pitch can be added to build up an ensemble
MDS:	Multi-Dimensional Scaling.
MIDI:	Musical Instrument Digital Interface.
Mixture:	A pipe organ stop that has multiple ranks, or sets, of pipes, designed to reinforce the upper harmonics of other pipes.
Pipe organ:	a complex musical instrument using wind-blown pipes to produce sound. See section 1.4.1 on page five for more details of this and other related glossary terms.
PCA:	Principal Component Analysis
Principal chorus:	an ensemble of pipe organ stops of the principal family, speaking at several harmonic multiples of the unison.
Reed:	A type of organ pipe that works like an oboe, producing a more harmonically complex sound than a flue pipe.
Specification:	the tonal scheme of an organ.
Spectral centroid:	the mid-point of spectral energy distribution on a frequency scale.
Spectral slope:	the average decrease in harmonics as they ascend in number.
Spectral smoothness:	the average absolute difference in amplitude between adjacent harmonics.
Stop:	A control that determines which pipes speak at any given time.
Timbral semantics:	the study of words relating to timbre.
Timbre:	an attribute of sounds distinct from their amplitude and pitch, which is intrinsically related to their frequency spectrum.
Unison:	A stop of 8' pitch. Middle C of a unison stop is the same frequency as Middle C of any other keyboard instrument in the same tuning and temperament.
Upperwork:	Stops more than an octave above the unison, such as mixtures.

References and Bibliography

Anderson, Poul-Gerhard, 1969. *Organ Building and Design*, George Allen and Unwin.

ANSI, 1960. USA Standard Acoustical Terminology (Including Mechanical Shock and Vibration), *Technical Report,* American National Standards Institute, New York. Report S1.1-1960 (Revised 1976).

Berger, Kenneth W., 1965. *Some Factors in the Recognition of Timbre*, Journal of the Acoustical Society of America, vol. 36, part II, pp. 1888-1891.

Bergweiler, S., Bergnerl, A., Görne, T., Wegener, M. and Gerhard-Multhaupt., R., 2003. *Breathing modes and sound radiation of metallic organ pipes*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol. 1, pp. 317-319.

Beurmann, Andreas and Schneider, Albrecht, 2003. *Sonological analysis of harpsichod sounds*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol. 1, pp. 167-170.

Bicknell, Stephen, 1996. *The History of the English Organ*, Cambridge University Press.

von Bismarck, G., 1974 (1). *Sharpness as an attribute of the Timbre of Steady Sounds*, Acustica, vol. 30, no. 3, pp. 159-172.

von Bismarck, G., 1974 (2). *Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes*, Acustica, vol. 30, no. 3, pp. 146-159.

Bonavia-Hunt, Noel, 1939. *The Organ of Tradition*, London, p16.

Bregman, Albert S., 1990. *Auditory Scene Analysis*, MIT Press.

Campbell, S. L., 1994. *Uni- and Multidimensional Identification of Rise Time, Spectral Slope, and Cutoff Frequency*, Journal of the Acoustical Society of America, vol. 96, no. 3, pp. 1380-1387.

Charboneau, G. R., 1981. *Timbre and the Perceptual Effects of Three Types of Data Reduction*, Computer Music Journal, vol. 5, no. 2, (Summer 1981), pp. 10-19.

Chowning, J., 1999. *Perceptual Fusion and Auditory Perspective*, in Music, Cognition, and Computerised Sound, ed. Cook, P. R., MIT Press, pp. 261-275.

Churcher, B. G., 1965. *Organ Tone and Acoustical Environment*, The Organ, vol. 45, pp. 68-85.

Churcher, B. G., 1970. *Some Researches on Organ Tone*, The Organ, vol. 49, pp. 122-144.

Comerford, Peter J., 1981. *The Bradford Musical Instrument Simulator*, IEE Proceedings, vol. 128, Part A (5), (July 1981) pp. 364-372.

Comerford, Peter J., 1987. *Further Development of the Bradford Musical Instrument Simulator*, IEE Proceedings, vol.134, Part A (10), (December 1987) pp. 799-806.

Comerford, Peter J., 1993. *Simulating an Organ with Additive Synthesis*, Computer Music Journal, vol. 17, no. 2, pp. 55-65.

Cook, Perry (editor), 1999. *Music, Cognition and Computerized Sound – an introduction to Psychoacoustics*, MIT Press.

Creasey, David P., 1998. *An Exploration of Sound Timbre using Perceptual and Time-Varying Frequency Spectrum Techniques*, DPhil thesis, The University of York.

Dannenbring, G. L. and Bregman, A. S., 1978. *Streaming vs. Fusion of Sinusoidal Components of Complex Tones*, Perception and Psychophysics, vol. 24, pp. 369-376.

Davies, J. B., 1978. *The Psychology of Music*, Hutchinson, p121.

Deutsch, Diana, 1999. *The Psychology of Music*, Academic Press, 2nd edition.

Disley, Alastair C., 1999. Acoustic Signatures of Pipe Organs, MEng dissertation, The University of York.

Disley, Alastair C., 2000. Acoustic Signatures of Pipe Organs II, MSc dissertation, The University of York.

Disley, Alastair C., 2003. *A Case in Point*, The Organ, issue 325 (August 2003), pp. 24-26.

Disley, Alastair C. and Howard, David M., 2003. *Timbral Semantics and the Pipe Organ*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol. 2, pp. 607-610.

Disley, Alastair C. and Howard, David M., 2004. *Spectral correlates of timbral semantics relating to the pipe organ*, Proceedings of the Baltic-Nordic Acoustics Meeting 2004.

Dodge, C. and Jerse, T. A., 1985. *Computer Music: Synthesis, Composition and Performance*, Schirmer Books.

Dowling, W. Jay, 1999. *The Development of Music Perception and Cognition*, in Deutsch (1999), pp. 603-625.

Eagleson, H. V., and Eagleson, O. W., 1947. *Identification of Musical Instruments*, Journal of the Acoustical Society of America, vol. 19, p. 338.

Ethington, R. and Punch, B., 1994. *SeaWave: A System for Musical Timbre Description*, Computer Music Journal, Vol. 18, No. 1, pp. 30-39.

Fletcher, H., Blackham, E. D., and Stratton, R., 1962. *Quality of Piano Tones*, Journal of the Acoustical Society of America, vol. 34, pp. 749-761.

Fletcher, H., Blackham, E. D., and Christensen, D. A., 1963. *Quality of Organ Tones*, Journal of the Acoustical Society of America, vol. 35, p. 314

Freedman, M. D., 1967. *Analysis of Musical Instrument Tones*, Journal of the Acoustical Society of America, vol. 41, no. 4, pp. 793-806.

Frobenius, W., and Ingerslev, F., 1947. *Some Measurements of the End-Corrections and Acoustic Spectra of cylindrical open Flue Organ Pipes*, Transactions of the Danish Academy of Technical Sciences, no. 1, part 3.

Gabrielsson, Alf, 1985. *Interplay between Analysis and Synthesis in studies of music performance and music experience*, Music Perception, Fall 1985, vol. 3, no. 1, pp. 59-86

Geissler, Paul, Martner, Otto, Zerbs, Carsten and Schleske, Martin, 2003. *Psychoacoustic investigations on the possibility of aurally identical violins*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol.1, pp. 59-62.

Goad, Pamela. J., and Keefe, Douglas. H., 1992. *Timbre Discrimination of Musical Instruments in a Concert Hall*, Music Perception, Fall 1992, vol. 10, no. 1, pp. 43-62.

Gold, B., and Rabiner, L., 1969. *Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain*, Journal of the Acoustical Society of America, vol. 46, pp. 442-448.

Goodwin, A. W., 1989. *An Acoustic Study of Individual Voices in Choral Blend*, Journal of Research in Singing, vol. 13, no. 1, pp. 25-34. Reprint of Goodwin (1980), same title, in the Journal of Research in Music Education, vol. 28, no. 2, pp. 119-128.

Gordon, J. W., 1984. *Perceptions of Attack Transients in Musical Tones*, PhD Thesis, Stanford University.

Green, D. M., 1973. *Minimum Integration Time*, in 'Basic Mechanisms in Hearing', edited by A. R. Moller, Academic Press.

Gregory, Andrew H., *Timbre and Auditory Streaming*, Music Perception, Winter 1994, vol. 12, no. 2, 161-174.

Grey, J. M., 1975. *An Explanation of Musical Timbre*, PhD thesis, unpublished, Stanford University.

Grey, J. M., 1977. Multidimensional perceptual scaling of musical timbres, Journal of the Acoustical Society of America, vol. 61, no. 5, pp. 1270-1277.

Grey, J. M. and Gordon, J. W., 1978. *Perceptual Effects of Spectral Modifications on Musical Timbres*, Journal of the Acoustical Society of America, vol. 63, part 2, pp. 1493-1500.

Guettler, K., Schoonderwaldt, E., and Askenfelt, A., 2003. *Bow speed or bowing position – which one influences spectrum the most?*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol.1, pp. 67-70.

Hall, Donald E., 1980. *Musical Acoustics: An Introduction*, Wadsworth Inc.

Hall, Donald E., 1991. *Musical Acoustics*, Brooks/Cole Publishing, 2nd edition.

Hall, Donald E., 1993. *Musical Dynamic Levels of Pipe Organ Sounds*, Music Perception, Summer 1993, vol. 10, no. 4, pp. 417-434.

Handel, S., 1989. *Listening: An Introduction to the Perception of Auditory Events*, MIT Press.

Hartmann, W. M., McAdams, S., and Smith, B. K., 1990. *Hearing a Mistuned Harmonic in an Otherwise Periodic Complex Tone*, Journal of the Acoustical Society of America, vol. 88, pp. 1712-1724.

von Helmholtz, H. L. F., 1877. *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, Dover Publications, New York. 2nd Ed. Translated by A.J.Ellis (1954).

Horning, J., 1998. *Recording the Organ*, The Diapason, February 1998 pp. 16-18 and March 1998 pp. 13-15.

Howard, David M. and Angus, James, 2001. *Acoustics and Psychoacoustics*, Focal Press, 2nd edition.

Howard, J. H., and Silverman, E. B., 1976. *A Multidimensional Scaling Analysis of Sixteen Complex Sounds.* Perception and Psychophysics, vol. 19 part 2, pp. 193-200.

Hutchinson, W., and Knopoff, L., 1978. *The Acoustic Component of Western Consonance*, Interface, vol. 7, pp. 1-29.

Jeans, J., 1961. *Science and Music*, Cambridge University Press.

Kendall, Roger A., and Carterette, Edward C., 1993 (1). *Verbal Attributes of Simultaneous Wind Instrument Timbres: I. von Bismarck's Adjectives*, Music Perception, Summer 1993, vol. 10, no. 4, pp. 445-468.

Kendall, Roger A., and Carterette, Edward C., 1993 (2). *Verbal Attributes of Simultaneous Wind Instrument Timbres: II. Adjectives Induced from Piston's 'Orchestration"*, Music Perception, Summer 1993, vol. 10, no. 4, pp. 469-502.

Kendall, Roger A., and Carterette, Edward C., *Perceptual Scaling of Wind Instrument Timbres*, Music Perception, Summer 1991, vol. 8, no. 4, pp. 369-404

Kitching, L. K., 1987. From Small Beginnings, International Computer Music Conference Proceedings, p270.

Kitching, L. K., and Comerford, P. J., 1988. *Some Clues In Instrument Recognition*, Proceedings of the Computers in Music Research Conference, 1988, pp. 121-124.

Kob, M., 2000. *Eigenmodes of a Flue Organ Pipe*, Acustica, vol. 86, no. 4, pp. 755-758.

Kostek, B., Zwan, P., and Dziubinski, M., 2003. *Musical Sound Parameters Revisited*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol. 2, pp. 623-626

Leman, M., Vermeulen, V., De Voogdt, L., Camurri, A., Mazzarino, B., and Volpe, G., 2003. *Relationships between musical audio, perceived qualities and motoric responses – a pilot study*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol.2, pp. 631-634.

Levitin, D. J., 1993. *Experimental Design in Psychoacoustic Research*, in Music, Cognition and Computerised Sound, ed. Cook, P. R., MIT Press, pp. 299-327.

Madison, G., and Merker, B., 2003. *Consistency in listener's ratings as a function of listening time*, Proceedings of the Stockholm Music Acoustic Conference 2003, pp. 639-642 vol. 2.

Manly, B. F. J., 1994. *Multivariate Statistical Methods: A Primer*, Chapman and Hall, London, 2nd Edition.

McAdams, S., and Bregman, A. S., 1979. *Hearing Musical Streams*, Computer Music Journal, vol. 3, no. 4, pp. 26-43.

McVicker, William R., 1987. *An Analytical Approach to Open Cylindrical Organ-pipe Scaling from a Historical Perspective, with Specific Reference to the Scaling Practices of Selected Organ-Builders*, PhD Thesis, unpublished, Durham University.

Mercer, D. M. A., 1953. The Physics of the Organ Flue Pipe, American Journal of Physics, vol. 21, no. 5, (May 1953) pp. 376-386.

Miklós, A., and Angster, J., 2000. *Properties of the Sound of Flue Organ Pipes*, Acustica, vol. 86, no. 4, pp. 611-622.

Miller, J. R. and Carterette, E. C., 1975. *Perceptual Space for Musical Structures*, Journal of the Acoustical Society of America, vol. 58, no. 3 (September 1975), pp. 711-720.

Moore, B. C. J., Glasberg, B. R., and Shailer, M. J., 1984. *Frequency and Intensity Difference Limens for Harmonics within Complex Tones*, Journal of the Acoustical Society of America, vol. 75, pp. 550-561.

Moore, B. C. J., Glasberg, B. R., and Peters, R. W., 1986. *Thresholds for Hearing Mistuned Partials as Separate Tones in Harmonic Complexes*, Journal of the Acoustical Society of America, vol. 80, pp. 479-483.

Moravec, Ondrej and Štěpánek, J., 2003. *Verbal description of musical sound timbre in Czech language*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol 2, pp. 643-645.

Nykänen, Arne and Johansson, Örjan, *Development of a language for specifying saxophone timbre*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol 2, pp. 647-650

Ohm, G. S., 1843. *Über die Definition des tones, nebst darangeknopfter Theorie der Sirene und Ähnlicker tonbeldener vorischtungen.* Ann. Deutsche Physics, vol. 48, pp. 497-565.

Owen, Barbara and Williams, Peter, 1988. *The Organ (New Grove)*, Macmillan.

Padgham, C., 1986. *The Scaling of the Timbre of the Pipe Organ*, Acustica, vol. 60, no. 3, pp. 189-204.

Pierce, J. R., 1983. *The Science of Musical Sound*, Scientific American Books, New York.

Piston, W., 1955. *Orchestration*, W. W. Norton & Co., New York.

Plomp, R., 1976. *Aspects of Tone Sensation: A Psychophysical Study*, Academic Press.

Pollard, H., and Jansson, E., 1982. *A tristimulus method of musical starting transients*, Acustica, vol. 51 pp. 162-171.

Pratt, R. L. and Doak, P. E.,1976. *A Subjective Rating Scale for Timbre*, Journal of Sound and Vibration, vol.43, no. 3, pp. 317-328.

Pries, A., 1984. *An Attempt to Describe the Parameter Determining the Timbre of Steady-State Harmonic Complex Tones*, Acustica, vol. 55, no 1, pp. 1-13.

Rasch, R. A., 1978. *The Perception of Simultaneous Notes such as in Polyphonic Music*, Acustica, vol. 40, pp. 21-33.

Rioux, Vincent, 2001. *Sound Quality of Flue Organ Pipes – An Interdisciplinary Study on the Art of Voicing*, PhD thesis, Chalmers University of Technology, Sweden.

Rioux, Vincent, 2000. *Methods for an Objective and Subjective Description of Starting Transients of some Flue Organ Pipes – Integrating the View of an Organ Builder*, Acustica, vol. 86, no. 4, pp. 634-641.

Rioux, Vincent and Västfäll, Daniel, 2001 (a). *Verbal Descriptions of Flue Organ Pipe Sounds (I): A Qualitative Analysis*, Musicae Scientae, also published in Rioux (2001), pp. 99-115.

Rioux, Vincent and Västfäll, Daniel, 2001 (b). *Verbal Descriptions of Flue Organ Pipe Sounds (II): A Quantitative Analysis*, Musicae Scientae, also published in Rioux (2001), pp. 119-115-132.

Risset, Jean-Claude and Matthews, Max V., 1969. *Analysis of Musical Instrument Tones.* Physics Today, vol. 22, no. 2, pp. 23-30.

Risset, Jean-Claude and Wessel, David L., 1999. *Exploration of Timbre by Analysis and Synthesis.* In Diana Deutsch, editor, *The Psychology of Music*, Academic Press, 2nd edition.

Rosch, E, 1975 (1). *Cognitive Representations of Semantic Categories*, Journal of Experimental Psychology: General, vol. 104, pp. 192-233.

Rosch, E, 1975 (2). *Cognitive Reference Points*, Cognitive Psychology, vol. 7, pp. 532-547.

Sandell, G. J., 1991. *Aconcurrent Timbres in Orchestration: A Perceptual Study of Factors Determining "Blend"*, PhD dissertation, School of Music, Northwestern University.

Scavone, G. P., Lakatos, S., Cook, P. R., and Harbke, C., 2001. *Perceptual Spaces for Sound Effects obtained with an Interactive Similarity Rating Program*, Proceedings of the International Symposium on Musical Acoustics, 2001, vol. 2, pp. 487-490.

Seashore, C. E., 1938. *Psychology of Music*, Dover Publications, New York. 1967 Reprint.

Shuter-Dyson, Rosamund, 1999. *Musical Ability*, in Deutsch (1999), pp. 627-651.

Solomon, L. N., 1958. *Semantic Approach to the Perception of Complex Sounds*, Journal of the Acoustic Society of America, vol. 30, no. 5, pp. 421-425.

Solomon, L. N., 1959. *Search for Physical Correlates to Psychological Dimensions of Sounds*, Journal of the Acoustical Society of America, vol. 31, no. 4, pp. 492-497.

Strang, G., 1969. *The Problem of Imperfection in Computer Music*, in 'Music by Computers', edited by H. von Foerster and J. W. Beauchamp, pub. Wiley, N.Y.

Sumner, William Leslie, 1962. *The Organ – Its Evolution, Principles of Construction and Use*, Macmillan.

Swannell, Julia (editor), 1986. *The Little Oxford Dictionary*, Clarendon Press, Sixth Edition.

Syrový, V., Otčenášek, Z., and Štěpánek, J., 2001. *Spectral Characteristics of Czech Baroque Pipe Organs*, Proceedings of the International Symposium on Musical Acoustics, 2001, vol. 2, pp. 477-480.

Syrový, V., Otčenášek, Z., and Štěpánek, J., 2003. *Subjective evaluation of organ pipe timbre in the standard listener positions*, Proceedings of the Stockholm Music Acoustic Conference 2003, vol.1, pp. 333-336.

Taylor, C. A., 1976. Sounds of Music, p49, BBC.

Thistlethwaite, Nicholas and Webber, Geoffrey, 1998. Editors, *The Cambridge Companion to the Organ*, Cambridge University Press

Van Bergeijk, W. A., Pierce, J. R., and David, E. E., 1960. *Waves and the Ear*, published by Doubleday, Garden City N.Y.

Von Békésy, G., 1960. Experiments in Hearing, McGraw Hill.

Von Bismarck, G., 1974. *Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes*, Acustica, vol. 30, pp. 146-159.

Index