Modelling and measurement of timbre perception in the electrically stimulated auditory system

by
Suzanne Hugo

Submitted in partial fulfilment of the requirements for the degree
Master of Engineering (Bio-Engineering)
in the
Faculty of Engineering, the Built Environment and Information Technology
UNIVERSITY OF PRETORIA

August 2010

SUMMARY

Modelling and measurement of timbre perception in the electrically stimulated auditory system

Author	:	Suzanne Hugo
Supervisor	:	Prof JJ Hanekom
Department	:	Electrical, Electronic and Computer Engineering
University	:	University of Pretoria
Degree	:	M.Eng (Bio-engineering)

SUMMARY

Music perception in cochlear implant (CI) listeners has been found to be generally unsatisfactory. An improved understanding of music perception in CIs is thus required, where research into the perception of timbre, an important aspect of music, could assist in improving this knowledge base. The aim of this study was to determine what underlies measured timbre perception in cochlear implantees. This was investigated by means of an experimental component and the development of a model of timbre perception in the electrically stimulated auditory system. Timbre perception was first measured in five normal-hearing (NH) and five CI listeners by means of three important timbre features, namely the spectral centroid, the logarithm of the rise time and the spectral irregularity. Discriminations of synthesised tones where these features were independently varied revealed that CI listeners had substantially larger threshold values than NH listeners for each of the timbre features investigated. Confusions of musical instrument timbres were also determined in five CI and five NH listeners by similarity ratings of original and acoustic simulations of musical instrument timbres, respectively. An acoustic model based on a six-channel advanced combination encoder (ACE) processor was developed in order to process 10 musical instrument timbres. The results of the similarity ratings revealed differences in the information conveyed through the timbre features for NH and CI listeners, and indicated that the acoustic model did not accurately

represent timbre in the electrically stimulated auditory system, but provided reasonable measurable results which could be compared to timbre perception model predictions. A model of timbre perception was developed by combining the results of the discrimination tasks with signal detection theory, in an attempt to predict the amount of information conveyed through each of the timbre features to both NH and CI listeners. The model was found to predict the experimental results obtained from the similarity ratings for both NH and CI listeners acceptably for each of the three timbre features. This outcome also confirmed the validity of the choice of the three timbre features as the primary features contributing to timbre perception, implying that timbre perception through a CI would be improved if CI processors could be optimised for the transmission of these three important timbre perception features. The model of timbre perception therefore has application in advancing CI research to facilitate music perception in the electrically stimulated auditory system.

KEY WORDS

timbre perception, cochlear implants, acoustic model, discrimination tasks, similarity ratings, model of timbre perception

OPSOMMING

Modellering en meting van timbrépersepsie in die elektries-gestimuleerde gehoorstelsel

Outeur	:	Suzanne Hugo
Studieleier	:	Prof JJ Hanekom
Departement	:	Elektriese, Elektroniese en Rekenaar-Ingenieurswese
Universiteit	:	Universiteit van Pretoria
Graad	:	M.Ing (Bio-Ingeniersewese)

OPSOMMING

Kogleêre implanting (KI)-luisteraars ervaar oor die algemeen onbevredigende musiekpersepsie. Beter begrip aangaande KI-luisteraars se musiekpersepsie vermoë, met spesifieke verwysing na timbrépersepsie, kan bydra tot die ontwikkeling van beter musiek prosessering stratgeieë. Hierdie studie het ten doel gehad om die onderliggende eienskappe van timbrépersepsie in KI-luisteraars te ondersoek. Ingesluit in die studie is ’n eksperimentele komponent, sowel as ’n modellerings-komponent om timbrépersepsie in die elektries gestimuleerde gehoorstelsel te simuleer. Drie belangrike eienskappe wat timbrépersepsie onderlê, naamlik spektrale swaartepunt, logaritme van die piektyd en spektrale onreëlmatigheid, is gebruik om die timbré van gesintetiseerde klanke te manipuleer. Tydens onderskeidingstake waaraan vyf normaalhorende (NH) en vyf KI-luisteraars deelgeneem het, het KI-luisteraars aansienlik hoër onderskeidingsdrempels as NH-luisteraars vir elk van die gemanipuleerde eienskappe getoon. Voorts is luisteraars se onderskeidingsvermoë ten opsigte van musiekinstrument timbrés ondersoek deur KI- en NH-luisteraars na onderskeidelik ware en gesimuleerde instrument timbrés te laat luister. Tien gesimuleerde musiekinstrument timbrés is geskep met behulp van ’n akoestiese model wat op ’n seskanaal ACE-prosesseerder gebaseer is. Eendersheidskattings het getoon dat die aard van die inligting wat KI-luisteraars ontvang het, nie soortgelyk was aan dié wat deur die gesimuleerde klanke aan die NH-luisteraars oorgedra is nie. Ten spyte daarvan dat die akoestiese model dus nie die omskakeling van

timbré-inligting soos deur ’n KI-prosesseerder bewerk, suksesvol benader het nie, het dit wel meetbare uitkomste daar gestel waarteen die uitsette van die timbrépersepsie-model vergelyk kon word. ’n Model van timbrépersepsie is geskep deur resultate van die onderskeidingstake met beginsels uit seindeteksie teorie te kombineer en sodoende die hoeveelheid inligting aangaande elk van die drie timbré eienskappe wat na die gehoorstelsel oorgedra word, te voorspel. Die model kon eendersheidskattings ten opsigte van die drie timbré eienskappe soos bepaal vir NH- en KI-luisteraars tot ’n aanvaarbare mate voorspel. Die bevinding bevestig dus dat die gekose eienskappe primêre bydraende eienskappe vir timbrépersepsie is. Indien KI prosesseerders dus beter oordrag van dié eienskappe daar kan stel, kan dit die weg baan vir verbeterde timbrépersepsievermoë van KI-luisteraars. Hierdie model van timbrépersepsie kan dus bydra tot navorsingspogings wat ten doel het om die musiekpersepsie vemoë van KI-luisteraars te bevorder.

SLEUTELWOORDE

timbrépersepsie, kogleêre implantings, akoestiese model, onderskeidingstake, eendersheidskattings, model van timbrépersepsie

LIST OF ABBREVIATIONS

ACE	:	Advanced combination encoder	(p. 29)
ADC	:	Analogue-to-digital converter	(p. 67)
ANOVA	:	Analysis of variance	(p. 111)
B	:	Brightness	(p. 59)
CI	:	Cochlear implant	(p. 2)
CIS	:	Continuous interleaved sampling	(p. 29)
eoa	:	End of attack	(p. 57)
eor	:	End of release	(p. 57)
FAT	:	Filter analysis table	(p. 67)
FFT	:	Fast Fourier transform	(p. 49)
FITA	:	Feature information transmission analysis	(p. 128)
GUI	:	Graphical user interface	(p. 103)
IIR	:	Infinite impulse response	(p. 72)
IDR	:	Input dynamic range	(p. 81)
IRR	:	Irregularity	(p. 59)
JND	:	Just noticeable difference	(p. 9)
LGF	:	Loudness growth function	(p. 68)
LRT	:	Log rise time	(p. 59)
MFCC	:	Mel-frequency cepstral coefficients	(p. 19)
MDS	:	Multidimensional scaling	(p. 17)
MLP	:	Mean logarithmic probability	(p. 138)
MPEAK	:	Multipeak	(p. 29)
NH	:	Normal-hearing	(p. 3)
NMT	:	Nucleus Matlab Toolbox	(p. 66)
PCA	:	Principal component analysis	(p. 19)
pdf	:	Probability density function	(p. 124)
pps	:	Pulses per second	(p. 34)
RMS	:	Root-mean-square	(p. 75)
SD	:	Standard deviation	(p. 27)
soa	:	Start of attack	(p. 57)
sor	:	Start of release	(p. 57)
SPEAK	:	Spectral peak	(p. 29)
SR	:	Stimulation rate	(p. 97)

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 PROBLEM STATEMENT

1.1.1 Context of the problem

The sensation of hearing is experienced when sound is presented to the auditory system, which consists of the outer ear, middle ear and inner ear. The primary mechanism in the process of hearing is the variation of air pressure, and the mechanical effect thereof on the tympanic membrane or ear drum (Fletcher and Rossing, 1998). The vibration of the tympanic membrane as a result of sound waves that pass along the ear canal causes the movement of a triplet of small, linked bones, which is then communicated to the fluid inside the spiralled cochlea. The pressure variations in the fluid of the cochlea result in the movement of the sensory hair cells that are in contact with the basilar membrane, inducing nerve impulses that are deciphered by the brain as the sense of sound (Clark, 2003; Fletcher and Rossing, 1998).

Sound has many forms and is an important part of life, whether as an essential part of communication, or for enjoyment purposes, such as listening to music. Most cultures across the world include music as a form of entertainment and art, which illustrates both its diversity and importance in life. Sound is a sensory experience that connects people in various facets of life, creating enriching experiences which would not be possible without the instrument of hearing.

Hearing loss may occur as a result of the functionality of the elements involved in the process of hearing being detrimentally affected. Examples of factors that may cause this include age, loud noise and certain drug treatments (Wilson and Dorman, 2008). Additionally, hearing loss can be caused by medical conditions, including environmental factors such as infections or head traumas, as well as genetic factors such as diseases (Willems, 2000; Loizou, 1999a). Conductive hearing losses refer to problems encountered with the outer ear and middle ear. Either a hearing aid or surgery can often assist with this type of hearing impairment (Fearn, 2001). Sensorineural hearing losses are primarily associated with a diminished number of hair cells in the inner ear (Clark, 2003), but can also refer to faults that hamper the neural communication of sound to the auditory temporal lobe (Fearn, 2001). A cochlear implant (CI) assists in restoring hearing by bypassing the hair cells that perform mechanical to neural translations, and instead stimulating the auditory nerves directly.

Generally, a CI consists of a microphone, a speech processor, a transmitter, a receiver and an electrode array. An ear level microphone transforms the sound into a waveform that can be interpreted by a speech processor worn on the body (Loizou, 1999a; Clark, 1996). The processor then encodes the sound into appropriate stimulus parameters which are transmitted inductively to the receiver (Clark, 2003). The receiver is placed under the skin and allows for communication with the external equipment (Wilson and Dorman, 2008). The receiver then directs the stimuli to the appropriate electrodes on the array positioned inside the cochlea, thus exciting nerve fibre populations and simulating the auditory nerve activity in response to sound in normal hearing, as discussed by Clark (1996).

CIs have developed rapidly over the past few decades, providing effective improvements in restored hearing to profoundly deaf people and enabling routine achievement of language perception in many candidates (Wilson and Dorman, 2008; Loizou, 1999b). Due to the success of CI development, research focus has recently shifted towards achieving perception in more difficult listening conditions, such as music (Pressnitzer, Bestel and Fraysse, 2005), in an attempt to provide CI users with advanced listening abilities and enjoyment (Lassaletta, Castro, Bastarrica, Pérez-Mora, Madero, de Sarriá and Gavilán, 2007).

Music and speech are both composed of complex structures of sounds and have many similarities (Limb, 2006). However, music is abstract and highly subjective in comparison to a spoken language, and this poses difficulties in defining and assessing the perception of music. Additionally, limitations imposed by CIs, such as the poor transmission of spectral information (McDermott, 2004; Kong, Cruz, Jones and Zeng, 2004; Pressnitzer *et al.*, 2005),

imply that musical characteristics may not be conveyed well or at all to the listener. The combination of the above factors illustrates the difficulties associated with understanding music perception in CIs.

Recently, research regarding music perception abilities in CI recipients has been carried out. Examples include work by McDermott (2004), Gfeller, Olszewski, Rychener, Sena, Knutson, Witt and Macpherson (2005), Leal, Shin, Laborde, Calmels, Verges, Lugardon, Andrieu, Deguine and Fraysse (2003), Kong *et al.* (2004), Fearn (2001) and Limb (2006). Such studies may serve as an entry-point into the understanding and development of CIs that are suited for musical perception and enjoyment.

1.1.2 Research gap

Simple perceptual inadequacies of the music perception abilities of CI listeners have been highlighted by music perception studies such as those listed above, in which perceptual abilities and differences in rhythm, pitch and timbre have been explained well for CI listeners. However, the results of such studies are often not quantitative, given the multidimensionality and subjectivity of music and, as a result, methods to overcome these perceptual inadequacies and improve the perception of musical sounds have not been sufficiently explored. To develop this area of research, quantitative results and conclusive methods to compensate for perceptual inadequacies in music perception for CI listeners are required.

Timbre has been highlighted as an important aspect of music, as it encompasses the characteristic quality of a sound (Risset and Wessel, 1982; Clarkson, Clifton and Perris, 1988) and includes the perceptual effects of a wide range of properties of acoustic signals. Therefore, research into timbre perception has the potential to facilitate an improved understanding of overall music perception in CIs. The existing knowledge of timbre and timbre perception in normal-hearing (NH) listeners may be applied to the case of the electrically stimulated auditory system as a basis for understanding timbre perception in CIs.

Characteristics of CIs can be investigated using methods such as psychoacoustic and speech recognition experiments with CI recipients. In terms of speech perception research, these methods can be effective and provide quantitative results. However, music is complex and difficult to define and this makes it very difficult for listeners, and in particular CI listeners, to explain what is heard from a piece of music. This highlights the need for a quantitative definition of music perception, especially in CI recipients, and a modelling approach could be the solution.

Models have been applied successfully in CI research to provide a quantitative understanding of speech perception in implant recipients. Examples include acoustic modelling, a method of representing acoustically what a CI recipient may hear as a result of electrical stimulation (Blamey, Dowell, Tong and Clark, 1984; Clark, 2003), allowing individual factors that affect auditory performance to be investigated without the complications of aspects such as subject variability and period of deafness. As a result, improved understandings of implant characteristics, leading to improvements in speech processing techniques for speech perception in CIs, have been achieved through acoustic modelling (Clark, 2003). Acoustic models have been applied effectively in speech perception improvements of CIs, but have yet to be applied for the purpose of advancing music perception in CIs. An acoustic model applied in this way would potentially achieve the same insight and advances in CIs suited to music perception as acoustic modelling has already achieved for speech perception. The implementation of an acoustic model in this study may provide a tool with which to test how changes in the processing may influence timbre perception, before this is tested directly on implantees.

Additionally, models of speech perception in CI listeners have been developed (e.g. Svirsky, 2000) and can be of great value in the stages of speech processor development preceding testing with cochlear implantees. Such a modelling approach applied to music perception in CIs may be the key to gaining a quantitative understanding of music as perceived through an implant. The development of a timbre perception model to provide a quantitative description of how listeners perceive timbre would provide insight into how a listener makes use of timbre information to perceive the timbre of the sound that was heard, allowing specific hypotheses regarding timbre perception to be tested in a rigorous manner. Specifically, a model that predicts the outcomes of timbre perception experiments correctly from features deemed important for timbre perception would allow conclusions to be drawn as to whether or not such features are in fact the primary features from which timbre is perceived, and whether or not CI listeners use these same features (to the extent that they are available) to perceive timbre.

1.2 RESEARCH OBJECTIVE AND QUESTIONS

Based on the discussion in the previous section, the primary objective of this study was to gain quantitative insight into the abilities of CI users to perceive timbre by developing a model of timbre perception to predict such abilities. To develop this idea, it is essential to understand the important characteristics of timbre. This understanding can be facilitated by studies that define timbre by a number of dimensions or features that are important for the correct perception of a musical sound (Grey, 1977; Krimphoff, McAdams and Winsberg, 1994; McAdams, Winsberg, Donnadieu, De Soete and Krimphoff, 1995). Important timbre features are also explained in studies pertaining to models of timbre, such as the model developed by Jensen (1999b).

To utilise existing timbre research to better understand timbre perception in CI users, an acoustic model approach is the most intuitive and would provide insight into the effects of the processing of a sound through a CI on the timbre of the sound. Important timbre perception features can be extracted from both the original sounds and the sounds processed through the acoustic model, with the aim of providing a quantitative representation of how CI listeners perceive timbre. Obtaining quantitative results of timbre perception using an acoustic model as a foundation, and in conjunction with psychoacoustic experiments, would then be possible. This could further be expanded on by developing a model of timbre perception, which could predict timbre perception abilities of both NH and CI listeners.

To facilitate the achievement of this main objective, the following tasks were set up to be accomplished.

- The decomposition of the timbre of a musical sound into quantitative features must be implemented, based on an existing model of timbre and knowledge of important timbre attributes.
- Measurements of timbre perception features in both NH and CI listeners must be performed to gain insight into the perceptual abilities of listeners and to provide a platform on which to develop a model of timbre perception.
- An acoustic model must be developed, based on existing acoustic models, through which musical instrument sounds can be processed to sufficiently represent timbre as conveyed to CI listeners.

- The timbre perception features extracted from both unprocessed musical instrument sounds and instrument sounds processed through the acoustic model, in conjunction with specifically formulated psychoacoustic experiments, should be used in the development of a model of timbre perception, to adequately predict the outcomes of timbre perception experiments for both NH and CI listeners.
- The model predictions should be compared to experimental results obtained from timbre perception studies performed with both NH and CI listeners, in order to draw conclusions regarding the validity of the acoustic model and the timbre perception model predictions.

The following main research question could be formulated and addressed to achieve the objectives of this study.

- Is it possible to develop a model of timbre perception that quantitatively represents timbre perception in both NH and CI listeners, that adequately predicts how a listener perceives musical instrument timbres?

To achieve the main objective of the study, several smaller objectives needed to be accomplished, as listed in the following points.

- Is it possible to correctly define and extract the most important timbre features from original and processed musical instrument sounds, to be used as a basis for predicting timbre perception?
- Can the acoustic model adequately predict which features important for timbre perception are transmitted to the auditory system of a cochlear implantee?
- Can quantitative conclusions be drawn as to how well important timbre features are conveyed to CI listeners?
- Can quantitative conclusions be drawn as to the differences in timbre perception for NH and CI listeners?
- Can these quantitative findings be adequately implemented in the development of a model of timbre perception that sufficiently predicts the outcomes of timbre perception experiments for both NH and CI listeners?

From the objectives and research questions posed it was intended that a quantitative description of timbre perception, forming a foundation for music perception, in CI recipients could be determined.

1.3 HYPOTHESIS AND APPROACH

To address the research questions that were posed in the previous section, it is important to have a model that defines the core characteristics of timbre according to measurable features. The effect of electrical stimulation on these features could then provide insight into timbre perception abilities of CI listeners.

Extensive research studies regarding timbre have been carried out over the past few decades, for example, those by Grey (1977) and McAdams *et al.* (1995), which have facilitated the development of a model of timbre by Jensen (2001). This timbre model encompasses what was found from the psychoacoustic experimental results in literature to be the most important characteristics of timbre. Using this model as a basis, an investigation into the timbre characteristics conveyed to the electrically stimulated auditory system was a possibility, with the ultimate aim of developing a model to quantitatively predict timbre perception abilities. Figure 1.1 illustrates the approach that was followed in this study to achieve this, with a description provided in the paragraphs that follow.

The important timbre features were extracted from the original instrument sound by using the methods of the existing timbre model by Jensen (1999b) and Jensen (2001), and timbre perception research by Krimphoff *et al.* (1994) and McAdams *et al.* (1995). Important timbre features that could be extracted are the spectral envelope, from which the perceptual feature of brightness can be extracted, the frequency envelope, from which the perceptual attribute of inharmonicity can be found, and the amplitude envelope, which is a substantial factor in discriminating between different musical instruments. These features can be used directly as inputs to the existing timbre model to resynthesise the sound (Jensen, 2001). This sound will be very similar to the original sound, thus representing the timbre features that are conveyed to a NH listener.

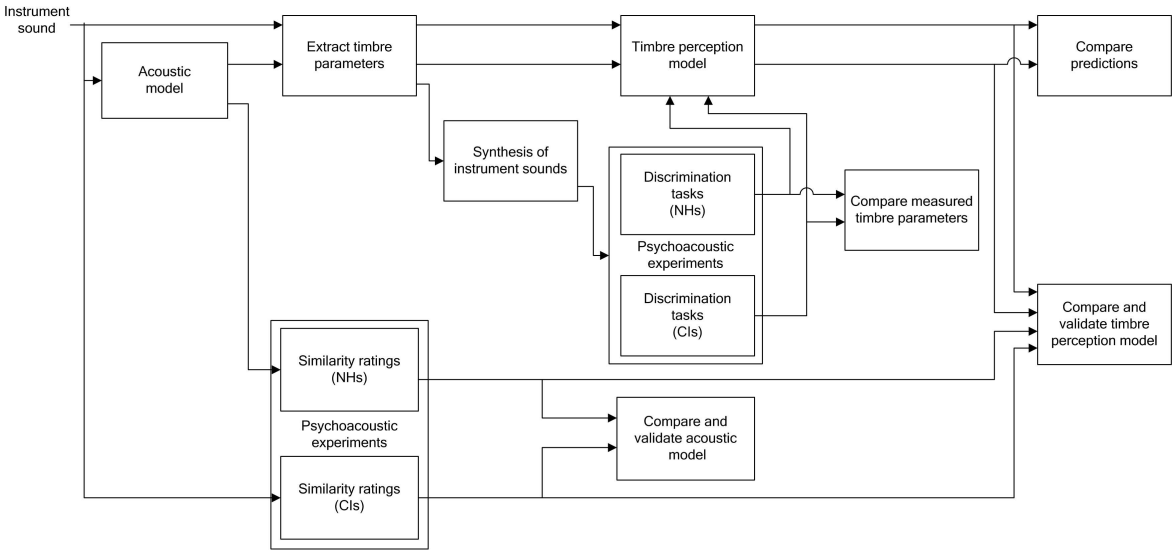

Figure 1.1.
Block diagram of the basic approach that was followed in this study for the representation of timbre in the electrically stimulated auditory system.

An acoustic model was developed, through which musical instrument sounds can be processed according to the effect of electrical stimulation. The model was based on existing acoustic models, some of which were presented by Clark (2003), and incorporated effects such as the limitations imposed by the number of electrodes as well as the speed at which a CI samples sound. These aspects are discussed in detail in section 2.4.3. The output of the acoustic model provided an estimation of what a cochlear implantee hears when listening to an instrument sound.

The use of an acoustic model is beneficial in that it allows for a large number of variations to be made in the model parameters, assisting in establishing the different effects on the recognition of a sound. Parameters can also be adjusted individually, which allows the effect of each parameter to be investigated independently.

The repeatability of acoustic models is an appealing feature, because factors such as the subjectivity of different CI listeners is ruled out. The variability of parameters is easier to control in NH listeners, without the complication of factors such as electrode placement in cochlear implantees. Additionally, there are more NH listeners available than CI listeners, indicating that experiments carried out with acoustic models can be conducted in larger quantities and in a much shorter time than experiments with implantees.

Using the output of the acoustic model, modified timbre perception feature values were extracted from the processed sound, by using the same existing timbre models and timbre feature calculations as those applied to the original musical instrument sound. This approach facilitated a quantitative comparison of the timbre features for CIs with the timbre model developed for NH conditions, to assess how well important timbre features are conveyed to cochlear implantees.

The extraction of timbre features from a sound allowed for synthesised sounds to be constructed from these basic features. Discrimination tasks for variations in important timbre features of synthesised sounds for both NH and CI listeners provided measurable timbre perception results in the form of discrimination threshold or just noticeable difference (JND) measurements, and comparisons between the abilities of the listener groups could then be made. Additionally, the JND measurements could also be used as parameters for the timbre perception model, assisting in defining the range in which instrument sounds were likely to be confused.

The timbre perception model was based on signal detection theory methods as presented by Svirsky (2000), and on the model by van Zyl (2008). The timbre perception model utilises the JND results of the discrimination tasks for each of the important timbre features, as well as the values of the timbre features extracted from the original and processed sounds, to make predictions as to the level of confusion between each possible combination of sound pairs.

The formulated predictions of timbre perception resulting from the model were compared to psychoacoustic experiments in the form of similarity ratings of musical instrument timbres, to validate the model. With NH subjects listening to acoustically modelled sounds and CI subjects listening to unprocessed instrument sounds, a comparison of these results to the predictions made by the timbre perception model could be made. The development of a model of timbre perception is a promising approach to quantitatively understanding timbre perception, and music perception in general, in CI listeners.

1.4 RESEARCH CONTRIBUTION

The limited understanding of music perception in CI listeners provides an unexplored field of research. Both a simple understanding of the perceptual abilities of CI users and quantitative models of musical sounds that already exist indicate that much of the knowledge basis required to better understand music perception in cochlear implantees already exists. By combining these aspects, the development of methods to improve music perception in CIs is feasible. This study is an entry-point to achieving the above, by making the following contributions.

- The extraction of important timbre perception features from both original sounds and those processed through the acoustic model, in conjunction with measurements of important timbre perception features obtained from discrimination tasks, could provide quantitative descriptions of timbre perception in both NH and CI listeners.
- These quantitative results could be implemented in the development of a model of timbre perception that could adequately predict the outcomes of timbre perception experiments and reveal, in measurable terms, how well timbre is transmitted to CI listeners. As a result, this model would present the possibility of accelerating CI research, with factors such as model repeatability making it a favourable option.
- The quantitative findings extracted from this study could offer insight into how the timbre features should be compensated for so that they may be transmitted effectively through a CI. This could assist in future endeavours for developing CI processing strategies suited to music perception, in the hope of advancing CIs and providing successful communication of music to cochlear implantees.

The content of each chapter is described briefly in the paragraphs that follow.

Chapter 2 presents the background necessary for understanding the research problem. As an introduction, definitions of music and some of its components are presented, with the focus on timbre. Following this, an overview of perceptual timbre research is given, a large part of which is based on MDS techniques. Analysis of musical sounds, with emphasis on the modelling of timbres of musical sounds, are discussed next. A background on CIs makes up the remaining sections of this chapter, providing a general literature study on the processing strategies employed in CIs, as well as acoustic model implementations, taking into account various CI characteristics. Music processing techniques employed in CIs are discussed briefly, with research regarding music perception in cochlear implantees concluding the literature study. Focus is placed on timbre perception research in CI listeners.

Chapter 3 discusses the methods employed in this study. This chapter is broken down into a number of main sections, each providing a description of the method followed to implement the particular aspect of the study. The first part introduces the musical instrument sounds that are used in the study. The relevant aspects of timbre modelling employed are described next, followed by a description of the important timbre features that form the basis of this study. The CI aspects of the study are discussed next, with the development of the acoustic model implemented explained in detail. This chapter presents the methods that form a basis for this study, on which the experimental and modelling components can be formulated.

Chapter 4 describes the experimental component of this study in detail. A large part of the content of this chapter has been submitted to Ear and Hearing in the form of a journal article for review and possible publication. This chapter thus includes specific methods, results and discussion sections for the experimental component of this study. Quantitative measurements of timbre perception for both NH and CI listeners are presented, obtained from psychoacoustic experiments. The experimental component consisted of two experiments, in the form of discriminations of timbre perception features and similarity ratings of musical instrument timbres, the results of which are discussed fully in chapter 4. Comparisons between these experimental results and the predictions of the timbre perception model can then be made, and are presented in chapter 5.

Chapter 5 describes the modelling component of this study in detail. The content of this chapter forms part of an article which will be submitted to Ear and Hearing as a continuation of the experimental findings of the article relating to the work of chapter 4. Chapter 5 thus also includes specific methods, results and discussion sections relating to the modelling work performed in this study. The development of the model of timbre perception, along with the analysis techniques used to assist in interpreting the model results are presented in this chapter. A full discussion of the developed timbre perception model is given, by comparing the predictions to the experimental results as well as to existing literature.

Chapter 6 presents a general discussion and conclusion of the study. The main discussion points from chapters 4 and 5 are summarised to provide a general discussion of the outcomes of the study with respect to the research questions posed in section 1.2 of this chapter. Following this, the main findings of the study overall are summarised. A critical analysis of the study is also presented, encompassing the implications of the study and directives for future research which may expand on the work in this study.

In summary, this chapter has provided an introduction to the work that will be addressed in this dissertation. A contextual background on hearing and CIs has been provided to present a framework for the study and to highlight the research gaps that will be addressed. In addition, a more focussed description of the research objectives and questions tackled in this study have been presented, as well as the basic approach followed to achieve these objectives. A contribution of the research to the field of CI technology has also be discussed, providing an indication of what this study aims to accomplish.

CHAPTER 2

LITERATURE STUDY

2.1 CHAPTER OBJECTIVES

To address the problem of quantitatively assessing timbre perception in the electrically stimulated auditory system as described in the previous chapter, an overview of existing related literature is required. This chapter presents and discusses the relevant literature required as a basis for this study and consists of a number of sections. Firstly, a brief introduction to music and music perception is given in section 2.2, followed by a more thorough background on timbre in section 2.3. This information serves as an entry-point to understanding music and timbre, as well as the perception thereof. Analysis of musical sounds as well as models of timbre are also be discussed, to gain insight into the physical characteristics and features of timbre. Section 2.4 provides an overview of CIs, describing the processing strategies that are relevant to this study, as well as a background on acoustic modelling to enable an understanding of sounds as processed through the electrically stimulated auditory system. Lastly, a background on music and timbre perception in CI listeners is presented, to provide an overview of the research findings on which this study is based.

2.2 MUSIC

To gain a better understanding of timbre, a brief background regarding music is required. Definitions of music and each of its components are necessary, and will be discussed briefly in section 2.2.1, followed by factors that may affect music perception and musical enjoyment in listeners, emphasising the subjectivity of music, in section 2.2.2.

2.2.1 Elements of music

Music is made up of complex structures of sound that can be produced by either instruments, or voice, or combinations thereof (Bregman, 2001; Clark, 2003). Music has many similarities with spoken language (Patel, 2003). Simply described, both speech and music employ sounds of varying frequencies presented over periods of time to convey a message (Limb, 2006; Bregman, 2001; Jensen and Marentakis, 2001), with the goal of communication and expression. The message conveyed can be either concrete, as in the case of speech, or abstract, as in the case of music (Donnelly and Limb, 2009). Music consists of several elements, regardless of genre or type (Limb, 2006), and can be basically categorised into pitch, rhythm and timbre (Clark, 2003), as discussed briefly in the following paragraphs.

Rhythm describes the temporal patterns in musical sounds with the time scale perception thereof usually in the order of seconds to minutes (Limb, 2006; Donnelly and Limb, 2009). As described by Ross (2008), rhythm is the grouping of a number of beats or steady sound pulses to create any series of durations of sound which may compel people to clap their hands or tap their toes in time. Temporal patterns that give rise to a distinctive rhythm occur in the approximate frequency range of 0.2 - 20 Hz (McDermott, 2004). Higher frequency components of acoustic signals convey pitch information.

Pitch describes the frequency of a musical sound, perceived as a note in a musical scale (Limb, 2006). A series of pitches that are structured into different musical contours and intervals form a melody (Clark, 2003; Donnelly and Limb, 2009; Ross, 2008). The perception of melody is very subjective, as any series of pitches that creates a sense of organisation or unity may be described as a melody (Ross, 2008).

Timbre is formally defined by the American Standards Association (1960) as “that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar”. Simply put, timbre is the quality that allows the distinction between two sounds with the same pitch, loudness and duration (Jensen and Marentakis, 2001), making the identification of several different musical instruments played simultaneously possible (Clark, 2003; McDermott, 2004). As discussed by Limb (2006), timbre results from spectral and temporal envelopes of a sound that interact in a complicated manner, allowing music to be defined as sounds of varying timbres that are organised in terms of rhythm, pitch and harmony (Limb, 2006). Timbre is involved in both the recognition of a familiar voice and the identification of a musical instrument (Donnelly and Limb, 2009) and is a complex and important element of music and sound in general (Krumhansl, 1989).

Harmony and counterpoint are additional important music components, as they consist of the basic elements of music. Harmony occurs when more than one pitch is played simultaneously and allows for the differentiation of the qualities of superimposed sound. This gives rise to other musical features such as counterpoint, which is a combination of multiple unfolding melodies in a musical piece (Limb, 2006; Clark, 2003). Most musical traditions have rules for the combination of notes that sound pleasant, known as consonance, or unpleasant, known as dissonance (Clark, 2003).

2.2.2 Factors that influence music perception

Although music can be structured into basic elements, music is perceived as a whole, which suggests that all aspects of music are equally important, as opposed to speech, in which there is a great deal of redundancy (Clark, 2003). The aim of research in music perception is to explain how a listener responds subjectively to musical sound signals, as stated by Rasch and Plomp (1982). Musical psychoacoustics are concerned with the relationship between the objective, physical properties of auditory stimuli in our environment, and the subjective, psychological responses evoked by them. In addition to the elements of music discussed in section 2.2.1, a number of subjective factors that influence music perception can be identified, and are discussed briefly in the following paragraphs.

Music is ultimately abstract and its interpretation is subjective, depending on a variety of factors. These factors include dynamics such as musical training and musical background,

which may heighten sensitivity towards a musical piece. Music listening habits and musical tastes, where the appraisal of musical genres may vary amongst different listeners, also reflect the diversity of musical listening experiences (Lassaletta *et al.*, 2007). This indicates that even cultural backgrounds may even affect the resulting perception of music.

The perception of music involves complex brain functions, as discussed by Koelsch and Siebel (2005), potentially affecting emotion and influencing the nervous, hormonal and immune systems. This suggests that personal emotional backgrounds and experiences are also likely to affect the resulting perception of music (Donnelly and Limb, 2009).

2.3 TIMBRE

The important timbre attributes extracted from perceptual findings as well as from physical models of musical instruments have led to the development of the timbre model by Jensen (2001). To model musical instrument sounds appropriately, a sufficient understanding of timbre perception is necessary (Jensen, 2002b). This section places the timbre model that will be used as a basis for this study into context, by providing an overview of the literature regarding timbre perception in section 2.3.1. The analysis of musical sounds is also reviewed in section 2.3.2, as the derivation of the timbre model results from an analysis by synthesis approach. From the literature on perceptual timbre research and the method of analysis of timbre by synthesis, important features of timbre have been extracted in order to implement the timbre model by Jensen (2001), which is highlighted in section 2.3.3.

2.3.1 Perceptual timbre research

Timbre is an auditory attribute that has been inadequately understood from a psychophysical perspective in the past, having been considered vaguely as a complex and multidimensional perceptual parameter of sound (McAdams *et al.*, 1995). Definitions of timbre, such as the standard given in section 2.2.1, tend to define timbre by what it is not rather than what it is (Risset and Wessel, 1982; McAdams *et al.*, 1995) in comparison to other well defined perceptual attributes of music. The multidimensionality of timbre makes it impossible to measure timbre on a single scale, such as soft to loud as in loudness perception, or low to high as in pitch perception (Rasch and Plomp, 1982), and introduces the difficulty of establishing,

through experiments, the number of dimensions and features required to represent timbre (McAdams *et al.*, 1995). Only in the past few decades has an improved understanding of timbre begun to emerge, with a number of different techniques utilised to facilitate this.

In the past, many studies regarding timbre and timbre perception have utilised forms of multidimentional scaling (MDS) techniques. These methods are useful for the study of complex stimuli, of which the perceptual or psychophysical characteristics are inadequately understood (Lakatos, 2000; Grey, 1977), thus finding effective application in understanding timbre perception. The MDS approach involves musical sounds as the starting point from which perceptual distances are measured, in an attempt to formulate a representation or coordinate system that explains the MDS axes (Terasawa, Slaney and Berger, 2005). This is achieved by applying MDS to data obtained from listeners rating differences between pairs of sounds for a number of musical timbres (Jensen and Marentakis, 2001). These results are then used with the aim of creating a map between the physical aspects of a musical instrument sound and the perceptual representation of each timbre attribute to a listener (Lakatos, 2000). Generally, MDS techniques generate two or three perceptual dimensions that can be interpreted.

Early examples of timbre research using the dimensional research approach are presented by Plomp (1969), where it was found for steady state musical tones that the three-dimensional map of musical tone similarities that was obtained could be interpreted entirely in terms of the amplitude pattern of the harmonics. Reports in other literature have shown consistent findings regarding the primary factors, namely spectral information, that facilitate timbre perception. Wedin and Goude (1972) performed analyses on musical sounds with attack and decay portions included and found their structure of perceptual dimensions of musical instrument tones to have a clear correspondence to the spectral envelope properties of the sounds. Miller and Carterette (1975) used a set of defined timbre attributes to create synthetic sounds and varied temporal and spectral properties, namely the amplitude envelope and number of harmonics, respectively, as well as the onset patterns of the harmonics. They found that important factors for the perception of timbre similarities were the number of harmonics as well as the amplitude envelopes and onset rates of the harmonics, which suggests that spectral characteristics were dominant in the perception of timbre.

Grey (1977) developed a three-dimensional perceptual model of timbre, in which the first dimension related to the spectral energy distribution, and the other two dimensions related to a number of temporal patterns of the tones. These included synchronicity in the higher harmonic rise and decays and thus levels of spectral fluctuations, as well as the presence

of low-amplitude, high-frequency energy in the initial attack segments. Grey and Gordon (1978) experimented with the effect of spectral modifications on musical timbres and compared the MDS analysis of the modified sounds to the MDS analysis of the original sounds in Grey (1977). They found that the sounds that exchanged spectral energy shapes exchanged orders along the spatial axes acquired in Grey (1977), validating the interpretation of the perceptual space using MDS analyses.

Krimphoff *et al.* (1994) analysed three-dimensional spaces and found the centre of the sound spectrum, the logarithm of the rise time, and the spectral flux to be the important acoustic correlates. McAdams *et al.* (1995) illustrated a new MDS technique to assign a large number of listeners with varying musical experience into a small number of underlying classes. Five class structures were found for a three dimensional spatial model, where musical training showed an ambiguous relation to this classification. The common dimensions of their model were quantified psychophysically in terms of the logarithm of the rise time, spectral centroid and degree of spectral variation or spectral flux. Lakatos (2000) attempted to better isolate the dimensions of timbre, generalised over a wide range of timbres and psychophysical techniques including MDS analyses. It was found that the spectral centroid and rise time alone adequately represented the most important perceptual dimensions of timbre, independent of musical training.

Studies by Samson, Zattore and Ramsay (1997) and Caclin, McAdams, Smith and Winsberg (2005) involved MDS analyses from the perception of synthesised tones by means of dissimilarity ratings of sound pairs. Samson *et al.* (1997) varied the spectral and temporal properties of their synthesised sounds and from MDS found that spectral information and rise time were the two independent perceptual dimensions that emerged, in accordance with studies mentioned previously. The study by Samson *et al.* (1997) included experiments with both single tones and melodies, and no distinct differences were noted between the two cases in defining the perceptual space. This indicates that enough information may be transmitted in single tones alone, and that the intricacies of a melody do not provide additional information in such a task.

Caclin *et al.* (2005) used synthesised sounds to vary the spectral centroid, rise time and other spectral properties deemed important features for timbre perception in past literature. Their findings indicate that the spectral centre of gravity, the logarithm of the rise time and the spectral fine structure or irregularity in the spectrum of the sound are the three most important dimensions in timbre perception.

Loureiro, de Paula and Yehia (2004) branched away from research involving comparisons of isolated notes of different musical instruments and focussed on the mapping of spectral characteristics of musical timbres produced by one instrument. A large variety of sounds produced by the clarinet were investigated by means of principal component analysis (PCA) techniques, to obtain a set of spectral bases or dimensions from which the different timbres could be categorised. It was shown that timbre classes were dependent on the spectral brightness of each sound.

Although MDS and related analyses have been the primary tool in forming an understanding of timbre perception, other methods have been implemented to achieve this. Examples include work by Terasawa *et al.* (2005) and Terasawa, Slaney and Berger (2006), in which, as opposed to MDS, a defined coordinate system is used as a basis from which different sounds are created according to this representation. Each sound representation is then measured to determine a fit to the defined perceptual space (Terasawa *et al.*, 2005). This method is known as the Mel-frequency cepstral coefficients (MFCC) model and is shown to be a good model of timbre perceptual space (Terasawa *et al.*, 2005; Terasawa *et al.*, 2006). These studies address the representation of timbre, but only in a static form. However, sound is not static and factors such as rise and decay times have been shown to be important in timbre percpetion, thus these works only form a basis on which to build a complete model of timbre (Terasawa *et al.*, 2006).

De Poli and Prandoni (1997) conducted a series of experiments in which they attempted to algorithmically develop timbre spaces from a defined experimental framework. The results exhibited similarities to past literature making use of MDS analysis, and showed potential in exploring timbre qualities through an analytical approach which would not require subjective ratings of listeners. Other methods to gain insight into timbre perception include spectral simplifications to establish the discrimination thresholds or JNDs of acoustic signal changes (e.g. Jensen and Marentakis, 2001).

Other focussed research regarding timbre perception by Clarkson *et al.* (1988) studied timbre perception in infants. This enabled an identification of the spectral cues that infants make use of, by presenting complex tones with spectral and temporal information selectively added. The results indicated that infants can analyse the spectra of complex tones and discriminate differences in the spectral envelopes, one of the most important cues in timbre perception.

The studies discussed in the above paragraphs each employ a type of perceptual test that may be used to evaluate timbre perception. These tests can be categorised into a few groups, including verbal attributes, where the listener has to describe a musical sound by means of words such as sharp or dull, full or empty and colourful or colourless (Jensen, 2002b). Additional test categories are dissimilarity and discrimination tests, as discussed by Jensen (2001) and Jensen (2002b), where judgements in differences between musical sounds are made. Dissimilarity tests involve analysing judgements made regarding differences between the timbres of two different musical instruments and discrimination tests analyse judgements in differentiating between original and modified musical timbres of the same instrument.

Bregman (2001) discusses auditory stream segregation as another method of examining the qualities of timbre in relation to the perception thereof. A sequence of sounds may be heard as either originating from a single source, such that it is perceived as one integrated stream, known as fusion, or as originating from distinct sources, such that it is perceived as two segregated audio streams, known as fission (Cooper and Roberts, 2007). Auditory stream segregation or auditory streaming is an occurrence in which a quick sequence of high and low tones separates into two distinct perceptual streams, one with the high tones and the other with the low tones (Dannenbring and Bregman, 1976; Chatterjee, Sarampalis and Oba, 2006).

In experiments by Singh and Bregman (1997), a distinct lowering effect on both the fission and fusion boundary fundamental frequency value could be noted when adjusting the timbre properties of the middle tone in a repeating three-tone sequence. It was shown that spectral differences in timbre were significant for stream segregation, whereas there was some debate as to whether or not temporal differences, for example, in the attack and decay of the timbres, were important in stream segregation. These results provided leads to important features of timbre that could be used as a basis on which to develop a model of timbre.

2.3.2 Analysis of musical sounds

In addition to the perceptually important timbre features that have been investigated in section 2.3.1, other physical aspects of timbre should also be considered, and form part of the timbre model implemented by Jensen (1999b) and Jensen (2001). One of the most well known methods of analysing the physically important features of sounds is by means of additive parameters which constitute a good analysis and synthesis model of voiced sounds

(Jensen, 2002b). In the paragraphs that follow, some physical attributes of timbre are noted, and a description of the analysis by synthesis approach of investigating the physical properties of musical instrument sounds is provided.

As described by Hartmann (2005), tone colour refers to the timbre of the steady state segment of a sound, that is, the part of the sound without onset and offset transients. This entails the part of the sound that is not related to sensations of loudness or pitch (Zwicker and Fastl, 1999). By this definition, it is necessary to extract from the mixture of sensations the features that may be relevant in recognising timbre. According to Zwicker and Fastl (1999) these may be qualities such as sharpness, or inversely pleasantness, which in turn depend on sensations such as tonalness and roughness. In the case of a pure tone, the tone colour depends only on frequency, i.e. a low frequency (below 200 Hz) will sound dull, while a high frequency (above 2000 Hz) will sound sharp or piercing. This indicates that it is the frequency content and not the shape of the waveform that determines tone colour. Risset and Wessel (1982) and Zwicker and Fastl (1999) confirm this by stating that sharpness relates to the spectral content, specifically the position of the spectral envelope along the frequency axis.

Most musical instrument sounds are composed of a fundamental tone and a number of harmonics. String, woodwind and brass musical instruments generally act as lowpass filters that attenuate harmonics with frequencies greater than 1000 Hz (Hartmann, 2005). The difference in timbre produced by different musical instruments, as described by Zwicker and Fastl (1999), is a result of the frequency spectra or relative amplitudes of their harmonics. For example, the flute produces mainly one frequency component (the fundamental frequency), while the trumpet produces a number of harmonic components and a broader frequency spectrum.

Rasch and Plomp (1982) discuss temporal characteristics, such as onset effects, as well as steady state effects, being important in the recognition of timbre. These may include factors such as the rise time and shape of the rise curve and the presence of noise during the onset times, as well as factors such as pitch instability over time. Jensen (1996) also discusses how the amplitude envelopes over time are affected by the control of musical instruments. For example, the envelope of a piano tone depends on the speed at which the note is played, and factors such as the decay depend on how long the note is held for, affecting the important temporal timbre features.

An additional relevant point is that acoustical instruments can be divided into two classes, envelope-based instruments, and continuous-control instruments (Jensen, 1996). Some instruments, such as bowed string instruments like the violin, are capable of both techniques, where plucking the string forms part of the envelope-based class and stroking the bow on the strings constitutes continuous-control.

Due to the complex physical behaviour of musical instruments, it is difficult to isolate specific fixed characteristics of musical instrument sounds. As discussed by Risset and Wessel (1982), this highlights the need to extract important features from a complex physical structure, which may be achieved by exploring timbre by means of analysis and synthesis. The analysis and synthesis of musical instruments are generally achieved by using a model of a sum of sinusoidals (Jensen, 1999b), known as an additive model. This method has also been implemented in speech analysis (McAulay and Quatieri, 1986), and is a practised and effective method of analysing sounds. Jensen (1999b) summarises the early techniques for analysing the additive parameters of musical sounds, dating back to more than a century ago, that provided insight into musical instrument tones.

More recent research regarding the study of musical instruments by analysis of additive parameters includes work by Ando and Yamaguchi (1993), in which a statistical study of the spectral parameters in musical instrument tones was performed. Here an initial decomposition of musical sounds into additive parameters was carried out, from which it was concluded that by incorporating the statistical properties of musical tones into the synthesis of sounds, a high sound quality would potentially be achieved.

The additive model approach is implemented in the timbre model (Jensen, 2001) used in this study. The additive model was chosen due to the existing knowledge regarding analysis/synthesis properties, with well understood parameters such as time, frequency and amplitude, as well as the perceptually expressive parameters of this model (Jensen, 2002b). As discussed by Jensen (2002b) and Jensen (2001), the additive analysis consists of associating a number of sinusoids with a sound, and estimating the time-varying amplitudes and frequencies of these sinusoids. The sound can then be resynthesised by summing the sinusoids to produce a highly realistic sound.

Jensen (1999b) explains that the sinusoids correspond to the harmonic overtones when the sound is harmonic, in which case the frequencies of the sinusoids are multiples of the fundamental frequency and are equally spaced in distance in the frequency domain. The first

number of extracted frequencies correspond closely to the notes in the 12 tones per octave scale and therefore the relationship between the frequencies of compound musical sounds determines the consonance of the musical interval (Kameoka and Kuriyagawa, 1969). The additive parameters can best be visualised in the form of a three-dimensional plot, as shown in figure 2.1, with axes corresponding to time, frequency, and amplitude. The lines in the plot, known as partials, indicate the time evolution of the amplitude and frequency of each sinusoid. As an example, a harmonic test signal with a fundamental frequency of 100 Hz is shown.

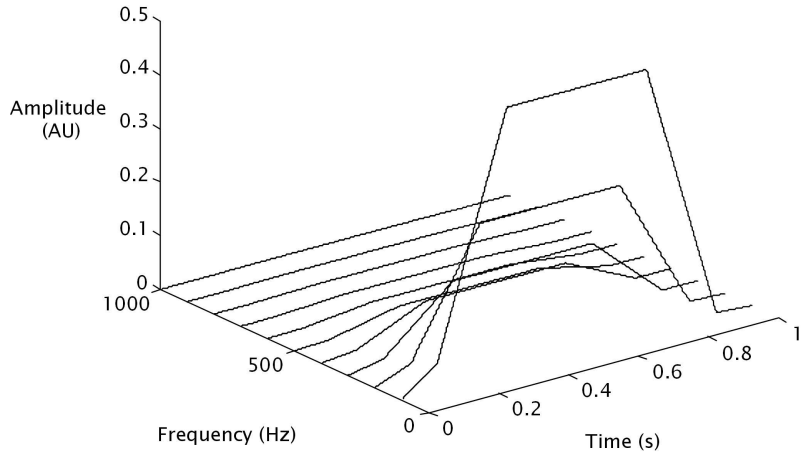

Figure 2.1.
Example of the additive parameters of a harmonic sound with a fundamental frequency of 100 Hz.

The closest line on the frequency axis in figure 2.1 is the fundamental frequency. The amplitude of the fundamental component is first zero for 100 ms, then it follows a linear rise for 200 ms, a plateau for 300 ms, then a linear decay for another 200 ms and returns to zero for the remainder of the one second length of the sound. A total of 10 partials are shown, with each component having an amplitude that is half the amplitude of the preceding partial.

2.3.3 Models of timbre

Most of the parameters of the timbre model (Jensen, 2001) have an intuitive perceptual quality due to their relation to timbre perception, and many of them can be related to the physics of musical instruments (Fletcher and Rossing, 1998). The timbre model by Jensen (1999b) and Jensen (2001) was inspired by perceptual research on timbre, as described in section 2.3.1, but was derived from the analysis of musical sounds using the method of analysis by synthesis, as described in section 2.3.2. Based on these research findings and methods, general conclusions can be made regarding the most important timbre features that are included in the timbre model. These conclusions are described in the following paragraphs, along with other timbre models that have been developed based on similar concepts.

Jensen (2002a) states that, in general, the most important timbre features can be extracted from the amplitudes and frequencies of a sound. These features are loudness, defined as the maximum of the amplitude in a log scale, and brightness, as extracted from the amplitudes, and the fundamental frequency and inharmonicity as extracted from the frequencies. In summary, the spectral envelope, temporal envelope and irregularities of a sound can be highlighted as the most important timbre features (Jensen and Marentakis, 2001; Jensen, 2001). The timbre model (Jensen, 1999b; Jensen, 2001) incorporates the most significant timbre attributes, as listed below, which will be elaborated on in the paragraphs that follow. The timbre attributes as given by Jensen (2001) that are incorporated into the timbre model are:

- the spectral envelope, associated with the brightness and resonances of the sound,
- the frequency envelope, associated with the pitch and inharmonicity of a sound,
- the amplitude envelope, consisting of five segments: start, attack, sustain, release and end, each segment with an individual start and end relative amplitude and time, and
- irregularities, separated into amplitude irregularities, known as shimmer, and frequency irregularities, known as jitter.

The spectral envelope has been found to be one of the most important timbre features (Grey, 1977; McAdams *et al.*, 1995). The shape of the frequency spectrum is the key to this attribute, as it shows the amount of energy present at each frequency across the audible range (Clark, 2003). The perceptual feature of brightness is associated with the centre of gravity

of the spectral envelope (Jensen and Marentakis, 2001; Marentakis and Jensen, 2001). Resonances occur as a result of the shape or structure of a musical instrument (Clark, 2003) and give rise to formants, the composition of which is important to the timbre of an instrument. It is noted by Clark (2003) that there is no significant change in the formant structure as the notes of an instrument change, indicating that all pitches played on a particular instrument have similar timbres.

The frequency envelope encompasses the simplicity of the behaviour of the harmonics over the course of the note (Bregman, 2001). Clark (2003) states that the steady-state frequency is an important component in an instrument's timbre as, for example, the frequency spectrum of a clarinet contains almost only odd-numbered harmonics. The frequency envelope models the deviation of each partial from the harmonic case (Marentakis and Jensen, 2001), which relates to the perceptual attribute of inharmonicity. In the case of a piano, for example, the stiffness of the strings causes the higher partial components to have much higher frequencies than in the harmonic case, and produces a degree of inharmonicity in the sound (Jensen, 1999b).

The amplitude envelope is important as there is a strong continuity of the frequency spectra over time for musical instruments (Clark, 2003). Figure 2.2 illustrates the different segments in a typical amplitude envelope model, using the fundamental partial of the test signal of figure 2.1.

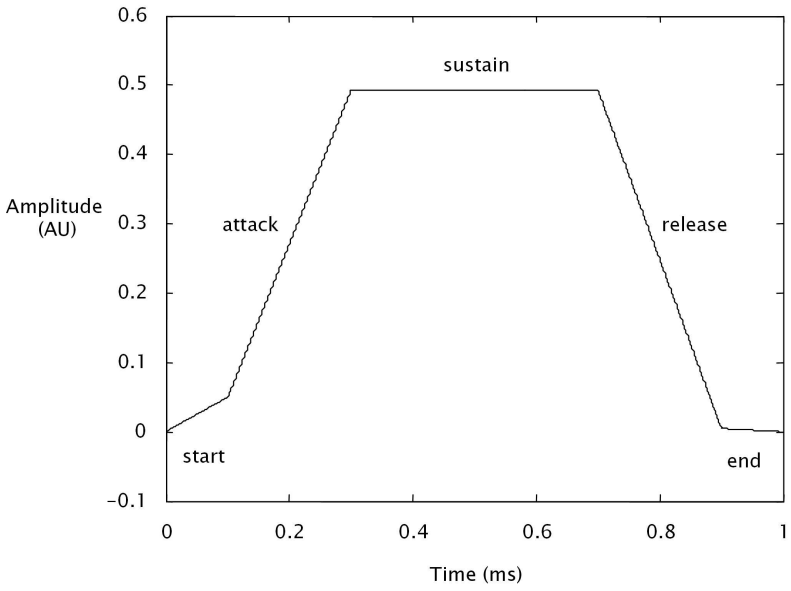

Figure 2.2.
Example of the five amplitude envelope segments for a musical instrument, as implemented in the timbre model by Jensen (1999b).

Bregman (2001) and McAdams *et al.* (1995) state that the attack or rise time is possibly the most important amplitude envelope parameter, with the time of this segment resulting in a specific slope, as well as the irregularities present in the attack segment playing an important role in the recognition of timbres. According to Marentakis and Jensen (2001), the slope of the sustain segment can be used to distinguish between instrument sounds that are played continuously, for example, that of the flute, from sounds that decay automatically, for example, that of the piano. In addition, the duration of the release segment allows a distinction to be made between damped and non-damped sounds.

The timbre model finds the start and end points of each of the five segments and then fits a curve between each of these points to represent an approximation of the amplitude envelope for each partial. Irregularities can then be added to the smooth approximated envelopes to represent the sound more accurately.

Irregularities in the sound (both shimmer and jitter) provide for slow random variations of the frequencies and amplitudes and also for additive noises that may occur in the instrument

sound (Jensen and Marentakis, 2001), which give a real quality to the sounds (Jensen, 2001). The timbre model by Jensen (1999b) includes a model for shimmer and jitter in the form of low pass filtered Gaussian noise with a given standard deviation (SD) and bandwidth. Marentakis and Jensen (2001) observe that the perceptual effect of altering the SD of the noise can be likened to altering the signal to noise ratio, which eventually results in the sound becoming unvoiced. The bandwidth of the noise is a way of adjusting the speed of the noise signal fluctuations. This can range from slow random variations to uncorrelated noise, depending on the filter bandwidth. Finally, there are the shimmer and jitter correlation groups, the purpose of which is to control how much the shimmer or jitter noise signal of each partial is correlated to the noise of the fundamental partial. The modelling of the noise parameters adds liveliness and a real quality to the sound.

The timbre model consists of a number of partials, of which the amplitudes are the sum of a clean envelope made up of an attack, sustain and release segment, with irregularity (shimmer) added. This amplitude is then multiplied with the spectral envelope value, where the frequency is the sum of a static value and irregularity (jitter). The timbre features used to model timbre can then be described as:

- the maximum amplitudes of each partial,
- envelope model times, amplitudes and curve form coefficients for each partial for the attack, sustain and release segments,
- the mean frequencies, and
- irregularity SD, bandwidth and correlation.

Timbre models find application in the automatic classification of musical instruments. This is a difficult procedure to carry out, but has become the focus of research pertaining to computer music and grouping of music into genres. Attempts in this field of research have been made by Herrera-Boyer, Peeters and Dubnov (2003), using the concept of a model of timbre, such as the model by Jensen (2001), to achieve automatic musical grouping. Other models such as that of Bensa, Jensen and Kronland-Martinet (2004), have developed a detailed but very specific model of a realistic piano sound, which provides a good representation of the timbre of the piano, but lacks a generalised description encompassing many timbres. As asserted by Jensen and Marentakis (2001), the timbre model by Jensen (2001) models all voiced isolated

musical instruments and has an intuitive parameter set of a fixed size, which separates the sound into dimensions that relate to the timbre dimensions proposed in research, providing a promising, well established representation of timbre.

2.4 COCHLEAR IMPLANTS

With an understanding of timbre and the timbre model of Jensen (2001) in place, a background on CIs is required on which to base an adequate representation of timbre perception in cochlear implantees. A background on CI processing strategies will be presented in section 2.4.1, focussing on strategies used in the implementation of the acoustic model in this work. The factors that may influence the performance of CIs are presented next, followed by an overview of acoustic modelling, which will be used to provide a CI-mediated representation of timbre. Following this general background, a more specific overview of music processing in CIs will be given, as well as factors that may influence music perception in implant recipients.

2.4.1 Processing strategies

The speech processor plays an important role in the success of CIs because it is responsible for deriving the most appropriate stimuli to be presented to the electrodes. Many signal processing strategies have been developed over the years, as discussed in detail by Loizou (1999b) and Clark (2003). In general, speech processing strategies involve dividing the input signal into a number of different channels in an attempt to present stimuli to the electrodes to effectively imitate the natural firing patterns inside the cochlea, thus optimising the speech intelligibility of the listener.

The frequency of a sound is encoded through both rate coding and place coding, as described by Clark (1996). Rate coding refers to the encoding of the frequency of sound by varying the rate of electrical stimulation on the cochlea. Place coding refers to the encoding of the frequency of sound using multiple electrodes placed according to the tonotopic organisation of frequencies in the cochlea. Initially, CIs were implemented using single-channel implants (House and Berliner, 1982), whereby electrical stimulation was provided at a single site in the cochlea using one electrode (Loizou, 1999b). With the development of CIs, multichannel

implants became the norm, to allow spectral information to be transmitted more readily by making use of the place mechanism for coding frequencies (Loizou, 1999a).

In multichannel CIs, different electrodes are stimulated depending on the frequency of the signal. High frequency signals cause stimulation of electrodes situated close to the base of the cochlea, while low frequency signals stimulate the electrodes situated close to the apex. The main function of a CI signal processor is to filter the input audio signal into different frequency bands or channels, which can then be applied to the corresponding electrodes, in an attempt to mimic the function of a healthy cochlea (Loizou, 1998). Many design considerations in the development of CIs arise as a result of multichannel implants. These include the number of channels that are sufficient for adequate speech understanding (Friesen, Shannon, Baskent and Wang, 2001; Dorman, Loizou and Rainey, 1997b), as well as the type of information that should be transmitted to each electrode (Whitford, Seligman, Blamey, McDermott and Patrick, 1993). To address these design considerations, different signal processing devices with varying numbers of spectral channels have been developed. Generally, a CI consists of a fixed number of implanted electrodes with a selection of these electrodes activated depending on the number of spectral channels of the implemented processor.

The many different types of signal processing strategies for CIs can be classified into three main groups, namely: waveform strategies, feature extraction strategies and hybrid strategies, as discussed by Loizou (1999b). In waveform strategies, such as the continuous interleaved sampling (CIS) approach, a waveform is derived from filtering the speech signal and presented as the stimuli. In feature extraction strategies, such as the F0/F1/F2 and multipeak (MPEAK) strategies, spectral features such as formants are derived using algorithms and presented as the stimuli. In hybrid strategies, such as “n-of-m” strategies, both waveform and feature extraction aspects are included and presented as the stimuli (Loizou, 1999b; Clark, 2003).

In “n-of-m” strategies in CIs, the speech signal is separated into m frequency bands and envelope information is derived from each band (Nogueira, Büchner, Lenarz and Edler, 2005; Loizou, 1999b). In traditional “n-of-m” strategies such as spectral peak (SPEAK) and the advanced combination encoder (ACE) strategies, the n envelope outputs with the largest energy out of the m bands are selected for stimulation (Loizou, 1999b). These strategies aim to neglect the less important features of speech and concentrate only on the significant spectral components to increase temporal resolution (Nogueira *et al.*, 2005).

The spectral maxima sound processor (SMSP) was the first technique to branch away from formant extraction techniques (Fearn, 2001) and form an “n-of-m” or peak picking strategy. It consists of 16 band pass filters analysed at a rate of 250 Hz. The commercial implementation of SMSP was expanded to form the SPEAK strategy, which includes 20 bandpass filters, as discussed by Whitford, Seligman, Everingham, Antognelli, Skok, Hollow, Plant, Gerin, Staller, McDermott, Gibson and Clark (1995). Once the maxima are selected, the corresponding electrodes are stimulated at an average rate of 250 Hz, but this may vary in the range of 100 Hz. The ACE processing strategy is an extension of the SPEAK strategy, with the stimuli either presented at higher rates or with more channels (Clark, 2003; Fearn, 2001).

As outlined by Wilson (2006), the ACE strategy generally makes use of a linear distribution of frequencies up to approximately 1300 Hz, after which a logarithmic distribution of frequencies is used, ranging up to the maximum frequency. In typical fittings of ACE processors, the number of electrodes, *m*, ranges from 20 to 22 and the number of activated channels, *n*, ranges from six to 16, depending on the implementation (Skinner, Holden, Whitford, Plant, Psarros and Holden, 2002). The maximum rate of stimulation with an ACE processor is 14400 Hz (Clark, 2003). In general, speech perception scores have indicated better results with the ACE processing strategy than with the CIS strategy, but did not show significantly different results from the SPEAK processing strategy (Skinner *et al.*, 2002; Clark, 2003).

2.4.2 Factors influencing the performance of cochlear implants

Given the many different processing strategies implemented in CIs, as well as subject variability, there is a great deal of inconsistency in the performance of CI recipients. Loizou (1999a) discusses some of the factors responsible for the variability of auditory performance, which are briefly outlined in the following paragraph.

Factors such as the duration of deafness of a subject prior to receiving a CI, relating to the age at which the onset of deafness occurred, generally affect auditory performance. For example, prelingual deafness will affect the learning of speech and language, as opposed to postlingual deafness, with detrimental effects on auditory performance. The duration of CI use may also affect auditory performance. Additional factors that affect auditory performance include the electrode placement and insertion depth of the electrode array inside the cochlea and the type of signal processing strategy employed (Loizou, 1998).

2.4.3 Acoustic modelling

The inconsistency of auditory performance among CI recipients makes it difficult to assess the various factors that affect speech perception. These factors may also not be independent of one another, heightening the difficulty of assessing auditory performance factors individually (Loizou, 1999a). To address this problem, acoustic simulations are of great assistance as they represent acoustically what a CI recipient may hear as a result of electrical stimulation (Clark, 2003). This allows individual factors that affect auditory performance to be investigated without the complications of aspects such as subject variability and period of deafness.

An acoustic model may include different parts to emulate the effect of a CI on a sound, which can be separated into the processing part of the CI and the part of the CI that simulates the biophysical characteristics of the electrode-neural interface. To gain insight into the factors that influence the performance of CIs, acoustic models such as the model developed by Blamey *et al.* (1984), may be implemented. Generally, acoustic simulations involve the processing of speech in a similar fashion to a CI processor, whereby the speech signal is first filtered into different frequency bands or channels, which are used to stimulate the different electrodes spaced along the array inside the cochlea. However, in the case of an acoustic model, the output is presented acoustically as a sum of noise bands or a sum of sinusoids to NH listeners. Acoustic models have provided quantitative insight into speech perception, and even music perception in more recent research, such as that of Rubinstein and Turner (2003), in which the interaction between the number of spectral bands and the amount of temporal fine structure conveyed within each band was assessed. The results suggested that CI processing strategies that improved the coding of temporal fine structure were likely to improve both speech perception, especially in noise, and music perception in CI listeners.

The development of an acoustic model requires that both the signal processing factors and biophysical characteristics of the electrode-neural interface of CIs be considered. The latter encompasses effects regarding the physical location of the electrode array inside the cochlea, and includes factors such as current spread and insertion depth. A discussion of some specific acoustic modelling aspects that are addressed in this study is given in the following paragraphs, along with the effect of each of these modelled factors on the auditory performance of CI recipients.

2.4.3.1 Number of channels

An important factor that can be included in an acoustic model is the number of channels. The number of channels refers to the number of areas that are stimulated in the cochlea, and affects the level of speech perception. An optimum number of independent channels required for high levels of speech understanding must be found, and may be facilitated by means of acoustic simulations. Studies such as those by Dorman *et al.* (1997b) and Friesen *et al.* (2001) have investigated the effect of the number of channels on speech recognition in CI recipients. Nie, Barco and Zeng (2006) found that increasing the number of electrodes from four to 12 generally improved speech recognition, specifically regarding closed-set vowel recognition and sentence recognition in quiet. Additionally, eight to 10 electrodes were found to be optimal for speech intelligibility in noise (Fishman, Shannon and Slattery, 1997). However, as the channels increased between seven, 10 and 20, no difference was found in speech performance.

General conclusions from studies such as those mentioned above indicate that between five and eight independent channels are needed for good speech recognition (Loizou, 1998; Fearn, 2001), and so should be the number of channels typically implemented in an acoustic model.

2.4.3.2 Insertion depth

The insertion depth of the electrode array substantially affects speech performance in CIs. As explained by Loizou (1999a), electrode arrays are typically only partially inserted into the cochlea, usually 22 - 30 mm deep. This creates a frequency mismatch between the analysis frequency and the stimulating frequency. For example, as described by Dorman, Loizou and Rainey (1997a), if an electrode array is only inserted 22 mm into the cochlea, the most apical electrode will lie close to the 800 Hz frequency area of the cochlea. However, a typical centre frequency of 250 Hz of the first filter in the CI processor will then be used to stimulate the 800 Hz area, indicating that an upwards shift in frequency will take place, affecting the perception of speech.

Acoustic simulation studies that have been carried out to investigate the effect of the insertion depth of the electrode array on the performance of CI recipients include those by Dorman *et al.* (1997a), Faulkner, Rosen and Stanton (2003) and Baskent and Shannon (2005). As

discussed by Dorman *et al.* (1997a) and Loizou (1999a), it was concluded that the insertion depth significantly affects speech perception, as insertion depths of 23 mm and lower generally result in very poor speech recognition. Studies by Baskent and Shannon (2005) and Faulkner *et al.* (2003) showed that better speech recognition results are obtained when acoustic frequency information is mapped onto the corresponding cochlea place, using the frequency-to-place equations found in Greenwood (1990).

2.4.3.3 Channel interactions

As explained by Vanpoucke, Zarowski and Peeters (2004), an electrode on an electrode array inside the cochlea should ideally only excite neural fibres located in the immediate area of the electrode. However, the current that is injected spreads through the cochlea, exciting fibres that may be situated at a distance from the stimulating electrode. This occurrence results in channel interactions, a limitation which causes the number of perceptually independent frequency channels to be lower than the number of available electrodes (Friesen *et al.*, 2001).

Many factors can influence channel interactions in CIs, including the electrode configuration, e.g. bipolar or monopolar, as well as the placement and design of the electrode array, e.g. the distance between the electrode and the nerve cells (Bingabr, Espinoza-Varas and Loizou, 2008; White, Merzenich and Gardi, 1984). Vanpoucke *et al.* (2004) developed a model for an approximation of the current spread as a function of distance through the cochlea for each electrode stimulated separately. They found the current spread to be very wide and not strongly dependent on the place of stimulation. Throckmorton and Collins (2002) conducted an extensive study on the effect of channel interactions on speech recognition. It was found that various channel interactions, simulated by means of pitch reversals, forward masking and electrode discrimination, affect speech recognition to different degrees. In general, spectral interactions degraded speech recognition more than temporal interactions. The spectral interactions affecting lower-frequency information also caused a more detrimental effect on speech recognition than those affecting higher-frequency information.

Channel interactions may occur where neural populations associated with different electrodes overlap, with the degree of overlapping varying from subject to subject (Throckmorton and Collins, 2002). Research by White *et al.* (1984) presents a number of possible methods for reducing channel interactions, including the use of asynchronous stimulation as well as bipolar electrodes instead of monopolar electrodes. The use of biphasic pulses also stimu-

lates a smaller group of nerve cells, reducing channel interactions. Fu and Galvin III (2001) developed a model to desynchronise channels by introducing different delays for each channel. The study indicated that in the case of CI recipients where fine spectral structures are not available, cross-channel asynchrony in speech signals can be overcome by an increased spectral resolution. A recent study by Bingabr *et al.* (2008) showed the development of a new simulation for the effect of spread of excitation in CIs. Models such as these find strong applications in acoustic modelling, where the effect of the spread of excitation is included.

2.4.3.4 Rate of stimulation

There has been much debate regarding the optimal rate at which the electrodes of a CI should be stimulated. Vandali, Whitford, Plant and Clark (2000) studied the effect of different stimulation rates on speech recognition. The study found that higher stimulation rates sometimes showed improvements in speech performance, but could also produce undesirable effects, indicating the subject-specificity of this factor. The study also found no differences in speech perception when the pulse rate varied between 250 and 1615 pulses per second (pps) per channel (Fearn, 2001). However, as discussed by Fearn (2001), contradictory findings were recorded by Loizou, Poroy and Dorman (2000b), who found that a higher pulse rate always resulted in a positive effect. Stimulation rates of 2100 pps/channel resulted in improved speech scores when compared with rates of 800 pps/channel. Holden, Skinner, Holden and Demorest (2002) also studied the effect of stimulation rates and found that group mean speech perception scores for sentences and phonemes in noise across periods of time were significantly higher for a higher stimulation rate of 1800 pps/channel, compared to a stimulation rate of 720 pps/channel. Nie *et al.* (2006) found that increasing the rate from 1000 Hz to 4000 Hz for each electrode improved sentence recognition in quiet, but that this increase could degrade sentence recognition in the presence of competing voice. It was found that high-rate stimulation up to 2000 Hz is beneficial to speech perception, but an increase up to 4000 Hz may affect performance detrimentally due to the electrode interactions at this high rate.

2.4.3.5 Other factors

CI listeners (as opposed to NH listeners) have a limited dynamic range and spectral resolution. The large acoustic dynamic range of approximately 120 dB for NH listeners (Fu and Shannon, 1999) is compressed by a logarithmic function to as small an electrical dynamic range as 5 - 15 dB in CI processors (Loizou and Poroy, 2001). Loizou, Dorman and Fitzke (2000a) found that speech understanding in CI listeners could be severely impaired as a result of a reduced dynamic range, especially for vowel recognition. A study by Zeng, Grant, Niparko, Galvin, Shannon, Opie and Segel (2002) revealed that for optimal speech recognition in CIs, an input dynamic range of 50 - 60 dB is required. To accommodate this finding, a new amplitude mapping technique was presented by Zeng *et al.* (2002) to assist CI users with speech performance, where a logarithmic map is used for low frequency channels and a more compressed map is used for the higher frequency channels. Fu and Shannon (1998) investigated the effects of non-linear amplitude mapping in both CI users and NH listeners, concluding that inadequate amplitude mapping functions could cause the loudness growth to be unnatural, resulting in poor speech recognition. However, Fu and Shannon (1998) suggest that the application of simple logarithmic mapping functions could be sufficient to provide CI listeners with adequate speech recognition.

For high auditory performance, specifically for vowel identification, the spectral contrast, which is the difference between the spectral peak and the spectral valley, must be preserved to a certain degree (Loizou and Poroy, 2001). The spectral contrast is reduced in CI listeners mainly due to the reduced dynamic range, as well as due to amplitude compression. As discussed by Loizou and Poroy (2001), the steepness of the compression function used to map the amplitudes of the acoustic signal to the electric amplitudes presented to the electrodes is a contributing factor to reduced spectral contrast. Loizou and Poroy (2001) found that for high vowel recognition, CI listeners needed about 4 - 6 dB higher spectral contrast than NH listeners.

The factors discussed above that may affect the performance of CIs have also been jointly investigated. Loizou *et al.* (2000b) studied the effect of different CI processors on speech understanding by varying the parameters of the processors. It was found that the pulse rate and the pulse width had the most positive effect on speech recognition, where joint variations of these two parameters yielded higher speech performance in CIs. Other signal processing factors such as filter overlap and the shape of the amplitude mapping function were also investigated, but did not generate significant results.

Nie *et al.* (2006) investigated the contribution of spectral and temporal cues to CI speech perception. The effect of the number of electrodes, stimulation rate and temporal envelope extraction on speech perception in quiet and noise were evaluated. They found that a linear trade-off exists between the number of electrodes and the stimulation rate for consonant and sentence recognition in quiet, but not for vowel and sentence recognition in the presence of competing voice.

Acoustic models of CIs provide insight into the various parameters of CIs. The above-mentioned studies indicate that the limitations of CIs can be better located with acoustic simulations, allowing improvements in CIs to be more readily achieved. Acoustic models may also be used to test new CI design aspects (Rubinstein and Turner, 2003) to accelerate the development of CI technology.

2.4.4 Music processing in cochlear implants

Although speech and music differ, as described in section 2.2.1, they have structural similarities (Limb, 2006), introducing consequent difficulties in understanding speech and music processing independently. Recent studies by Peretz and Coltheart (2003) have shown that music processing in the central auditory system can be defined by a modular structure, by means of which differences and similarities between the modules used for speech and music processing are formulated. Essentially, the main differences appear to be in the different spectral and temporal requirements for music and speech (Zattore, Belin and Penhune, 2002; Zattore, 2001), indicating that the processing features required for music differ from those for speech. As discussed by Zattore (2001), Pretorius and Hanekom (2005) and Zattore *et al.* (2002), speech requires fine temporal information processing to which the left auditory cortex regions are better suited, and music requires fine spectral or tonal information processing, for which the right auditory cortex regions are specialised. This indicates that a processing system that can manage both temporal and spectral information with equal accuracy is required for correct speech and music perception (Pretorius and Hanekom, 2005).

Due to the inadequate spectral resolution of CIs, pitch processing capabilities appear to be the major downfall regarding music processing in CIs, because the perception of pitch changes, which essentially make up a melody, is drastically affected by the pitch processing resolution (Pretorius and Hanekom, 2005). The tonotopic organisation of the electrode array also contributes to poor pitch perception (Kong *et al.*, 2004), adding to the challenge of improving pitch resolution in cochlear implantees.

2.4.5 Music perception in cochlear implant recipients

Music perception abilities of cochlear implantees are still limited (McDermott, 2004), despite numerous research efforts in this field (Leal *et al.*, 2003; Gfeller *et al.*, 2005). In both speech and music perception tasks for cochlear implantees, the general approach is to perform psychoacoustic experiments, in which physical, measurable acoustic parameters are provided as inputs to the experiment and a subjectively based output is obtained from the listener's response to the task. This approach has been successful in measuring speech intelligibility in CI users, as the limited phonetic alphabets that exist in most languages provide listeners with a frame of reference through which sounds can be identified (van Wieringen and Wouters, 1999). This enables postlingually deafened listeners to distinguish between different speech sounds that are perceived, for example, different consonant sounds, providing distinct, conclusive information regarding the perception of specific speech components.

However, psychoacoustic experiments have revealed far less conclusive evidence regarding music perception abilities in cochlear implantees. This is mainly due to the fact that music perception cannot be measured in the same way that speech can from psychoacoustic experiments. Language perception is acquired from an early age, implying that the auditory system is trained to perceive speech (Shannon, 2005). Unlike language, music is not a necessity for communication and survival and is thus not developed as early or to the same degree as speech, as discussed by Limb (2006) and Shannon (2005). This creates difficulties in music perception tasks as listeners generally have an untrained musical ear. Furthermore, the fact that music is unrestricted in style and sound makes it difficult to measure exactly what is perceived by listeners.

In addition to these challenges in understanding music perception in cochlear implantees, music perception is also far more subjective than speech (Gfeller *et al.*, 2005; Limb, 2006), based on individual preferences and factors such as the listening habits of the CI user before and after receiving the implant (Gfeller, Woodworth, Robin, Witt and Knutson, 1997). Aspects such as memory of music would therefore play an important role in CI-mediated music perception. Lassaletta *et al.* (2007) and Gfeller, Christ, Knutson, Witt, Murray and Tyler (2000) discuss the fact that listening habits, including the number of hours spent listening to music and music enjoyment, decreased substantially after implantation. It was found, however, that more than half of the CI subjects still enjoyed listening to music following implantation. Additionally, numerous studies have shown that musical training for implantees can improve perception and enjoyment of musical listening, as noted by Donnelly and Limb (2009) and McDermott (2004). The above-mentioned subject-specific factors make it difficult to pinpoint the reason for the different perceptual capabilities regarding music in cochlear implantees.

Existing music perception studies, such as that of Gfeller *et al.* (2005), utilised real-world pieces of music, combining various elements of music such as rhythm, pitch and timbre, to measure perception of music in cochlear implantees, while other methods have focussed on separate elements of musical pieces (Galvin III, Fu and Nogaki, 2007; Gfeller, Witt, Woodworth, Mehr and Knutson, 2002c). Even though numerous experiments have been carried out, research has generally not provided measurable or conclusive results, but rather a generalisation of the abilities of cochlear implantees regarding music perception. For example, the addition of lyrics to a melody usually improves the performance of implantees in perception tasks (Gfeller *et al.*, 2005). However, it is still unclear which aspect of music perception enables this: the memory of the lyrics or of the speech processing capabilites of the implant (Pretorius and Hanekom, 2005).

However, according to Martin, Scheirer and Vercoe (1998), psychoacoustic studies provide great potential for better understanding musical content, as the limitations of music perception are highlighted as a result. Such studies provide useful resources from which important features that can be used in systems to understand music may be extracted.

General findings have concluded that rhythm is the attribute of music most readily perceived by CI listeners (Gfeller *et al.*, 1997; Leal *et al.*, 2003; McDermott, 2004). Studies such as those by Kong *et al.* (2004) and Leal *et al.* (2003) show that there is a correlation between performance scores in rhythmic tasks and speech perception tasks. The fine temporal resolution that is necessary for accurate speech perception in CI processors allows rhythm, which is made up of temporal components, to be perceived with higher performance levels.

Pitch and melody perception, however, are more challenging aspects for CI listeners (Limb, 2006), and without rhythmic cues, recognition of melodies may be severely impaired (Kong *et al.*, 2004; McDermott, 2004). Common methods of evaluating pitch perception consist of testing the recognition of familiar tunes or obtaining performance measures using simple pitch discrimination tasks (Pressnitzer *et al.*, 2005). Other studies, such as those carried out by Pijl and Schwarz (1995), used a single electrode stimulated by varying pulse rates and showed that temporal cues are capable of providing pitch information similar to NH subjects up to approximately 300 Hz (McKay, 2005). However, this approach uses a different technique from the normal process of sound transmission in the cochlea during acoustic sound perception (Limb, 2006). Pitch perception research in CI listeners has been performed by McDermott and McKay (1997), while melody perception studies have been carried out by Gfeller, Turner, Mehr, Woodworth, Fearn, Knutson, Witt and Stordahl (2002a) and Galvin III *et al.* (2007). Gfeller, Turner, Oleson, Zhang, Gantz, Froman and Olszewski (2007) provide a study that summarises how well CI recipients perform in music perception tests with different processing strategies, as well as with combined acoustic-electrical hearing compared to only electrically stimulated hearing.

Studies that have been carried out by Koelsch, Wittfoth, Wolf, Müller and Hahne (2004) indicate that similar potential brain response patterns occur in NH and CI listeners in detecting irregular-sounding musical sequences. These results suggest that the neural mechanisms to detect pitch and timbre relationships are active in implant users, implying that the pursuit of the improvement of music perception in CIs would be feasible, as the mechanisms to interpret music are present.

2.4.6 Timbre perception in cochlear implant listeners

In general, timbre perception is found to be unsatisfactory in CI users (Limb, 2006; McDermott, 2004), implying that in addition to pitch, timbre remains one of the more challenging aspects of music perception in cochlear implantees (Donnelly and Limb, 2009). Research on timbre perception in CI listeners can be separated into two main paths: timbre recognition and discrimination, and timbre appraisal, the subjective rating of the pleasantness of the timbre, as discussed in the following paragraphs.

2.4.6.1 Timbre recognition and discrimination

In general, studies on the perception of timbre in CI listeners have focused on the ability of listeners to either identify or discriminate different musical instrument sounds (McDermott, 2004). Examples of such studies include work by Leal *et al.* (2003) and reviews by McDermott (2004), resulting in findings that NH listeners regularly mistake musical instruments from the same family, such as different brass instruments (Donnelly and Limb, 2009). However, CI users show error patterns in identifying timbres that do not correspond to the type of instrument family (Donnelly and Limb, 2009), indicating poor timbre perception in cochlear implantees in general.

Gfeller, Knutson, Woodworth, Witt and DeBus (1998) studied timbre recognition and appraisal. Simple melodic patterns were played as solos on each of four musical instruments, namely the clarinet, the piano, the trumpet, and the violin. For timbre recognition, subjects were asked to identify the type of instrument producing the melody. The results showed that NH listeners recognised all of the instruments with a significantly higher accuracy than CI listeners. Errors in the recognition tasks of the NH listeners were most often within the same instrument family, while for CI listeners, the errors in recognition were more scattered.

A study by Gfeller *et al.* (2002c) showed that cochlear implantees found it more difficult to identify timbres when musical instruments were played in the higher frequency ranges than when those instruments were played in the lower frequency ranges. Cochlear implantees also found it more challenging to identify timbres from the family of string instruments (Gfeller *et al.*, 2002c), with percussion instruments the most readily identified (Limb, 2006). This again indicates that temporal cues are important in cochlear implantees for improved timbre perception (Donnelly and Limb, 2009).

General findings from the study by Gfeller *et al.* (2002c) revealed that under 50 % correct responses in identifying musical instruments were obtained by CI listeners, while NH listeners obtained more than 90 % correct responses. In support of this finding, a study by McDermott and Looi (2004), where subjects were asked to identify 16 different musical instruments, revealed similar findings. The results varied greatly across subjects as well as instrument types, with an approximate average of 44 % correct identification of all the musical instruments by the CI users and a significantly higher average of 97 % correct identification by the NH listeners (McDermott, 2004).

As discussed by Pressnitzer *et al.* (2005), the familiarity of the listener with the stimulus is essential for recognition tasks, implying that musical memory may have been measured unintentionally in the studies mentioned in the previous paragraphs. This again illustrates the inconclusive nature of the outcomes of timbre perception experiments for CI listeners.

In addition to timbre identification tasks, timbre perception in CIs has also been investigated by methods of forward masking (Stainsby, McDermott, McKay and Clark, 2002), where the perception of the steady-state envelopes of different musical instruments was examined. The shape of the internal spectrum was measured using forward masking, and in addition the ability of listeners to identify and discriminate between the same stimuli was also measured. Results showed that the strengths of the correlations of the better performing CI listeners compared well to NH listeners. This indicated that some CI users may have frequency selectivity that is comparable to that of NH listeners. Stainsby *et al.* (2002) also concluded that a large amount of spectral information seems to be available to CI listeners, which can be noted from their discrimination abilities. However, the performance in the identification experiments was poor, illustrating that steady-state spectral cues alone are not necessarily adequate to identify a musical instrument sound.

A recent study by Emiroglu and Kollmeier (2008) attempted to quantify differences in object separation and timbre discrimination between NH and hearing-impaired listeners. The experiments determined JNDs of timbre in NH and hearing-impaired subjects along continua of "morphed" musical instruments and investigated the variance of JND in silence and different background noise conditions and on different sound levels. Emiroglu and Kollmeier (2008) used the same database of sound recordings as those used in this study (described in section 3.2), but cut out the attack time for their stimuli. They investigated pairs of sounds that differed along three dimensions: spectral centroid, temporal flux, and a pair of sounds that varied in both temporal and spectral aspects. Morphing of the sounds was then carried out between these pairs, and JND values of the morphing parameter were investigated each time. A similar approach will be used in this study to investigate individual JNDs of perceptual features important for timbre perception.

2.4.6.2 Timbre appraisal

Timbre appraisal evaluations require that the listener describe the quality of musical instrument sounds to assess the pleasantness of a sound (McDermott, 2004; Gfeller *et al.*, 1998). This can be achieved by requesting the listener to assign either ratings, in terms of numbers, or adjectives, such as "clear" or "beautiful" to the sound quality. Gfeller and Lansing (1991) asked subjects to rate nine musical instruments to obtain descriptors of the perceived quality of musical instruments. The study took everyday life listening conditions into account in obtaining the quality ratings.

The timbre appraisal component of the study conducted by Gfeller *et al.* (1998) involved subjects rating different timbre samples on a sliding scale on the basis of how much they liked the sound. The resulting differences in appraisal between NH and CI listeners were substantial for two of the four instruments played, namely the trumpet and the violin, which were found to be far more pleasant to NH listeners.

Gfeller *et al.* (2002c) obtained measures of timbre appraisal when comparing CI listeners to NH listeners by means of numerical scales for overall pleasantness as well as for perceptual dimensions of dull-sharp, compact-scattered and full-empty. Average findings showed that the ratings of CI listeners were substantially lower than for the NH listeners, particularly in the ratings of string instruments.

Musical pieces from three genres of music, namely classical, country-western and pop, were presented to both CI and NH listeners in a study by Gfeller, Christ, Knutson, Witt and Mehr (2003), to rate the complexity and pleasantness of musical timbres. It was found that the CI users rated the musical excerpts to be more complex than did the NH listeners, with the least appraisal found for classical music.

In an attempt to improve the music perception abilities of CI listeners, researchers have explored the effects of training implant users. McDermott (2004) provides a summary of the training effects of CI recipients on music perception in general, while Gfeller, Witt, Adamek, Mehr, Rogers, Stordahl and Ringgenberg (2002b) focus specifically on the effects of training on timbre perception in CI recipients. The music training program used in the study by Gfeller *et al.* (2002b) was developed and described in detail by Gfeller, Witt, Kim, Adamek and Coffman (1999). In summary, the training program consists of 48 lessons (approximately 10 minutes of listening and responses per lesson) for a period of 12 weeks, and information regarding the families of musical instruments is incorporated. The results of the study by Gfeller *et al.* (2002b) showed that listeners that completed the training program showed significant improvements in their average timbre recognition and timbre appraisal scores when compared to the control group in which no improvements were recorded.

2.5 SUMMARY

Chapter 2 presented the literature on which this study was based. Using the existing timbre perception findings for both NH and CI listeners as discussed in this chapter as a foundation, experiments and models to assist with timbre perception measurements were developed. A foundation for the definition and extraction of important timbre perception features was provided, as well as methods of implementing acoustic models to represent sounds through the electrically stimulated auditory system. A summary of timbre perception literature as provided in this chapter enables the structured development of an approach to follow to achieve the objectives of this study. The methods followed to implement experimental procedures to measure timbre perception, as well as to develop a model of timbre perception, are discussed in detail in chapter 3.

CHAPTER 3

METHODS

3.1 CHAPTER OBJECTIVES

Using the background given in chapter 2 as a basis, the approach followed to quantitatively understand timbre perception in cochlear implantees and thus to be able to develop the model of timbre perception for electrically stimulated hearing is given in this chapter. The objective was to first implement fundamental parts of the existing timbre model of Jensen (1999b). This was done to enable the definition and extraction of important timbre features from original musical instrument sounds, as well as from acoustically modelled sounds, to investigate the effect of the electrically stimulated auditory system on the parameters of the timbre model. An acoustic model implemented to alter sounds according to the effect of electrical stimulation was developed in Matlab [1] version 2007b. The effect of electrical stimulation on the timbre features deemed important for timbre perception in NH conditions could then be used to predict the outcomes of timbre perception experiments for both NH and CI listeners by developing a model of timbre perception. This chapter presents the methods used to form a foundation on which to be able to address the research questions posed in chapter 1, leading up to the experimental and modelling components developed in this study, which will be discussed in detail in chapters 4 and 5.

[1] Matlab is a product of the MathWorks company (www.mathworks.com)

3.2 DATABASE OF MUSICAL INSTRUMENT SOUNDS

Ten different instrument sounds were used in the study to introduce a variety of musical timbres. These were obtained from the sound database of the University of Iowa (Fritts, No date) and were used throughout the study. The perceived loudness was very different across the stimuli presented in this database. Such vast differences in loudness would possibly have had a drastic effect on the similarity rating experiment results obtained. Thus, peak normalisation of the sounds was performed in an attempt to provide a more balanced perceptual level of loudness across the musical instrument sounds. Extensive details of the instrument sounds are given in section 4.2.2, as implemented in one of the experimental studies. These musical instruments included four main instruments, namely the piano, trumpet, clarinet and violin, as these can be played in a similar frequency range and each is a commonly recognised example of an instrument family (Gfeller *et al.*, 1998; Nimmons, Kang, Drennan, Longnion, Ruffin, Worman, Yueh and Rubinstein, 2008). These four musical instruments are used throughout sections 3.2 to 3.5 to illustrate the timbre parameters extracted and calculations performed on the musical instrument sounds. The note of each of these sounds was C4 (F0 = 262 Hz), and in each case the peak amplitudes of the sound were normalised. The sounds are illustrated in figure 3.1. In addition to these four primary musical instruments, six other musical instruments were included to encompass a range of musical timbres. The instruments were selected to include a variety of spectral and temporal properties, as well as representing more familiar musical instruments and their families (Galvin III, Fu and Oba, 2008).

The piano makes up the first family of instruments, namely pitched percussion or percussive string instruments (Gfeller *et al.*, 1998). The piano is the only sound of this family included in this study, as others (e.g. the harpsichord) are uncommon instruments and samples of the sounds are not readily available. Pitched percussion instruments are defined as having a string fixed at both ends as the primary source of vibration, with most of the energy radiated by the body of the instrument. Usually, all the frequency components (both even and odd) are present with inharmonic components prominently found (Strong and Plitnik, 1992). The attack or rise of the piano sound is very short, with a prominent key “thump” noise generated by string vibrations. The sustain part of the sound is brief or non-existent, and the fundamental component usually dominates the sound spectrum (Fletcher and Rossing, 1998).

The trumpet represents the brass family of instruments. Other instruments of the brass family included in this study were the French horn and the trombone. In these instruments, sound is produced by the vibration of the lips against a mouthpiece and along sections of cylindrical tubes (Gfeller *et al.*, 1998). The frequency components are again usually all present in brass instruments. For the trumpet specifically, a clearer and louder sound may be produced than in other brass instruments, changing the number of resonating frequencies in the tone (Strong and Plitnik, 1992).

The clarinet represents the woodwind family of instruments, with the flute and saxophone constituting the other two members of this group for this study. These instruments produce sound from oscillations in the air column as a result of a vibrating reed. In general, the odd frequency components of a clarinet are predominant up to around 2000 Hz, after which even and odd components are both present (Gfeller *et al.*, 1998). The violin represents the last family of instruments, the string family, with the cello and viola included as other examples of this group. The violin is the highest pitched instrument of the string family. Similarly to the pitched percussion instrument family, the primary vibrations in the string instrument family originate from a string fixed at both ends, with most of the energy produced by the body of the instrument, and to a smaller degree, by the string. All of the frequency components are usually present for string instruments, excepting those that contain a node at the point of excitation (Strong and Plitnik, 1992).

As discussed by Houtsma (1997), pitch is often confused with timbre and therefore, for the purposes of this study, all musical instrument sounds were played at the same pitch, chosen as Middle C or C4 (262 Hz). The octave surrounding and including C4 is the most common octave among the frequency ranges for western musical instruments, as discussed by Nimmons *et al.* (2008). F#3 (185 Hz) is the lower limit of the octave surrounding middle C (Nimmons *et al.*, 2008), with E4 (330 Hz) and G4 (391 Hz) being the other most common notes in familiar melodies such as nursery rhymes.

Examples of the original musical instrument sounds are illustrated for both the time and frequency domains in figures 3.1 and 3.2, respectively, for the four main families of instruments. These are represented by the piano, trumpet, clarinet and violin. The sounds are approximately two seconds in length, where in each case a single C4 note of the specific musical instrument is played. Illustrations of the other 6 instruments are given in Appendix A.

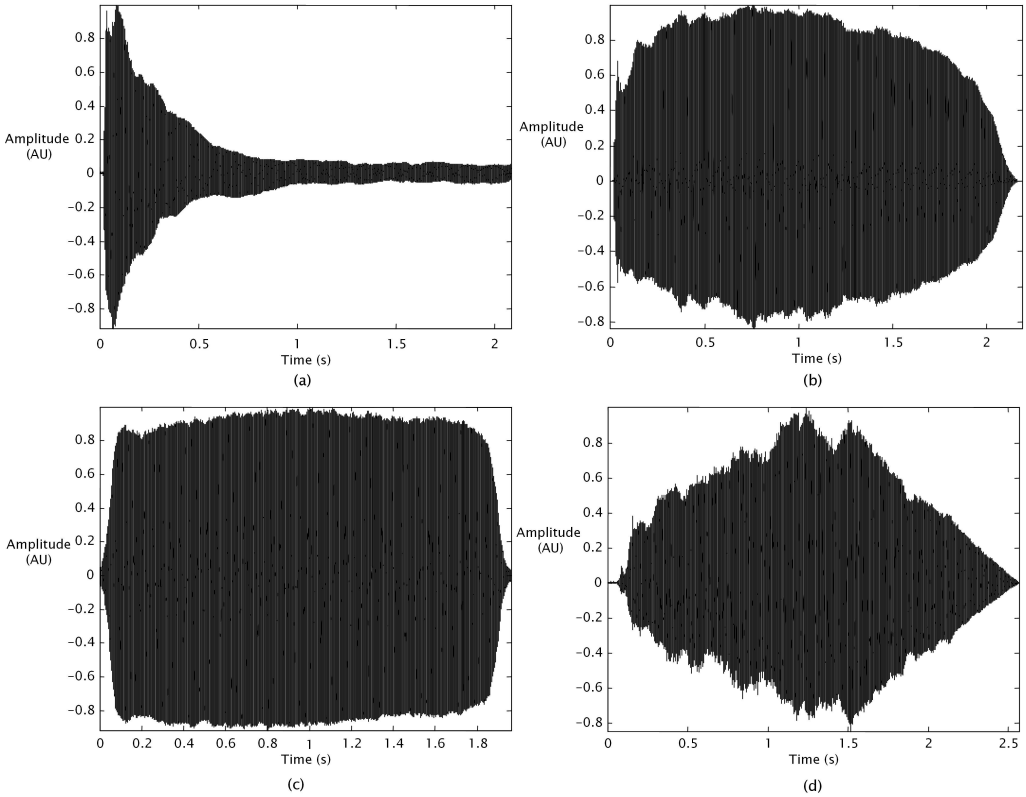

Figure 3.1.
Time domain representations of a selection of musical instrument sounds from each family of instruments with (a) the piano representing pitched percussion, (b) the trumpet representing brass, (c) the clarinet representing woodwinds, and (d) the violin representing the strings.

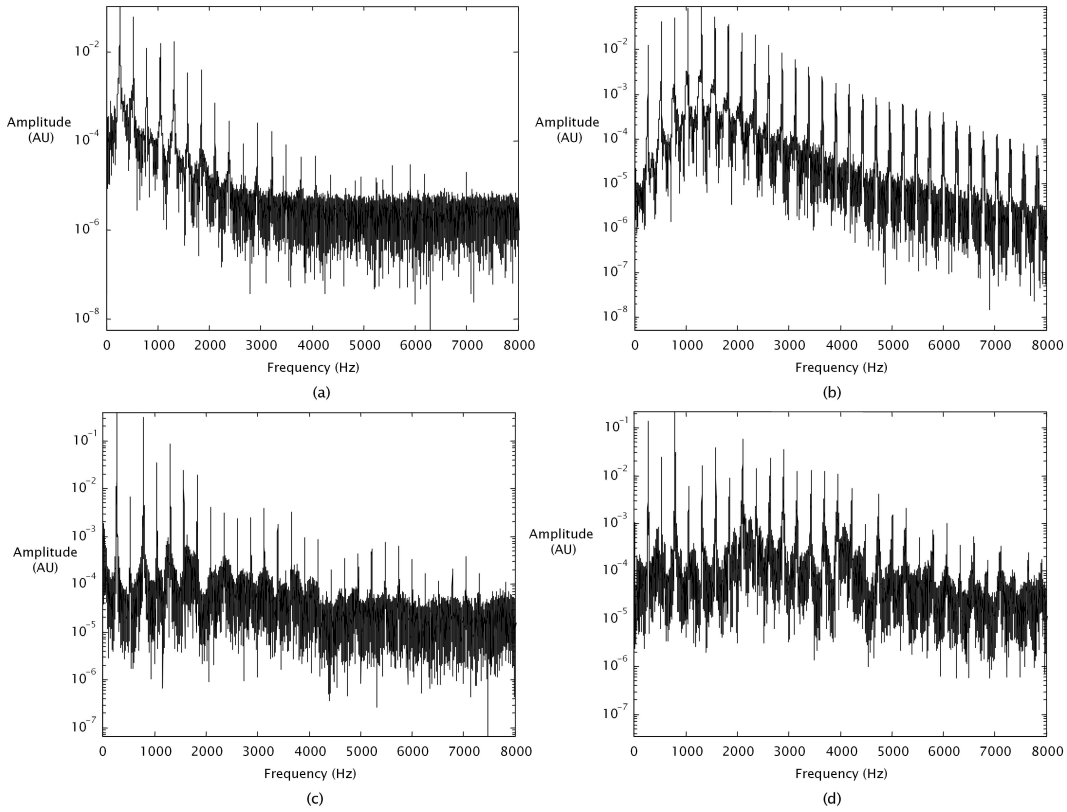

Figure 3.2.
Frequency domain representation of a selection of musical instrument sounds of (a) piano, (b) trumpet, (c) clarinet and (d) violin.

3.3 MODELLING TIMBRE

The implementation of the timbre model by Jensen (1999b) was used as a basis to extract various features of timbre to be used in the model of timbre perception developed in this study. The important steps in decomposing and analysing a musical instrument sound to extract the features that define the timbre are discussed in sections 3.3.1 to 3.3.5.

3.3.1 Fundamental frequency and frequency component estimation

The first step in analysing the timbre of a musical instrument sound is to decompose the sound into its frequency components. The original sound data are read from .wav files. The frequency analysis is performed over the strong segment of each sound (the entire sound after transient effects have been neglected). The fundamental frequency is generally defined as the first strong frequency component of a sound, or as the frequency difference between consecutive frequency components or overtones. In the timbre model by Jensen (1999b), these frequency differences are used to estimate the fundamental frequency. An additional improved method that fits the estimated frequencies to the ideal quasi-harmonic frequencies is also presented by Jensen to refine this procedure.

From the fast Fourier transform (FFT), the important frequency candidates of the sound are isolated by detecting the maximum points of the FFT. Using these candidates, the first estimation of the fundamental frequency can be made from equation 3.1, by calculating the mean of the differences between consecutive frequency candidates as

$$f_{\text{fund1}} = \frac{f_1 + \sum_{n=2}^{N} f_n - f_{n-1}}{N}, \tag{3.1}$$

where f_1 is the first frequency candidate and N is the total number of selected frequency candidates from which the first fundamental frequency estimation, f_{fund1}, is calculated.

To refine this estimation, frequency difference anomalies are removed by comparing frequency differences to the calculated fundamental frequency. If the differences between these frequency values exceed a certain threshold, the corresponding frequency difference value is discarded. This process is repeated, making the threshold smaller each time, until a desired small number, n, of frequency difference points are obtained. This gives fewer and more accurate frequency candidates which define the sound. The second refined fundamental frequency estimation can then be made by calculating the mean of the reduced frequency differences, again by using equation 3.1.

For some musical instrument sounds, for example the piano, the frequency difference of higher overtones can vary greatly from the fundamental. This characteristic is known as inharmonicity, and has been incorporated into the model of timbre by Jensen (1999b) to improve fundamental frequency estimation. To estimate the fundamental frequency, the frequency differences as described above, and here denoted as fd, are used. First, the differences between consecutive fd values are calculated and denoted as *fdd*. Next, the local average of the differences between *fdd* values over a few overtones is removed, shown in equation 3.2 as

$$fd_n^{'} = fd_n - \frac{\sum_{l=1}^{L} fdd_{n-l}}{L}, \tag{3.2}$$

where L is the number of overtones over which the local mean is calculated and removed, set as 3, giving the new frequency difference vector $fd^{'}$, taking into account inharmonicity. The improved fundamental frequency estimate can then be made by calculating the mean of $fd^{'}$.

With this fundamental frequency estimation, it is possible to recognise the frequency components, as found by the maximum peaks of the FFT, that are indeed harmonic components, as indicated by $fd^{'}$. Peak frequency values found to be harmonic values or close to harmonic values are retained. Frequency peaks that are not close enough in value to the harmonic components are eliminated.

It is also necessary to add harmonic components, as calculated from the fundamental frequency, that may be missing from the FFT analysis. To do so, the difference between the two overtone frequency values that precede the missing harmonic is calculated and added to the previous frequency component value to indicate the frequency at which the missing frequency component or overtone should be positioned.

Once all the overtone values have been included, the final fundamental frequency estimation can be made by fitting a stretched harmonic curve to the harmonic frequency points. The frequencies that are not exactly harmonic are said to be quasi-harmonic and can be expressed by the formula for a stiff piano string given by equation 3.3 as

$$f_k = k f_0 \sqrt{1 + \beta k^2}, \tag{3.3}$$

where f_k is the frequency for a specific overtone index, k, f_0 is the fundamental frequency and β is the inharmonicity value. By performing a non-linear least-squares curve fit to the harmonic frequency data, the values for f_0 and β can be found. The curve fitting is performed by the lsqcurvefit function in the Matlab Optimization Toolbox. To minimise errors in the important low partial components, the curve fitting is performed on the frequency components divided by the overtone index. Please refer to Jensen (1999b) for detailed documentation on the frequency estimation procedure.

3.3.2 Analysis of musical sounds by additive parameters

Once the fundamental frequency and frequency components have been obtained, the musical sounds can be analysed by means of an additive model. The musical sounds can then be modelled as a sum of sinusoidals constructed from the partial components of the sound, with time-varying amplitude and frequency, which when summed together resynthesise the sound with minimal loss of quality.

In this work, an FFT-based sliding time-domain window analysis is employed, whereby the FFT peaks are found by analysing the FFT of a windowed time signal. The peaks for a specific time segment are then attached to the partial tracks of the previous time segment. An optimum window length of four times the period of the fundamental is chosen (Jensen, 1999b), over which the FFT analysis for one time period is performed. The FFT is performed using a hamming window of the chosen length to avoid discontinuities.

Thus, the FFT of the sound signal multiplied by the hamming window is obtained for each time window. The window is shifted by 1/3 of the window length for each time interval and the FFT is calculated for each of these intervals. For each windowed FFT calculation, the maximum peaks that fall within a range of frequencies that correspond to the harmonic frequency components are calculated, as described in section 3.3.1. For each FFT peak located in this way, the respective frequency and amplitude values of these peaks are recorded for each time segment or window. This procedure gives rise to the time-varying amplitudes and frequencies for each frequency component, and allows the signal to be represented as a number of partials in time, frequency and amplitude.

The analysis described above results in the following representation in the form of additive parameters for each of the four instruments of figure 3.1, as shown in figure 3.3.

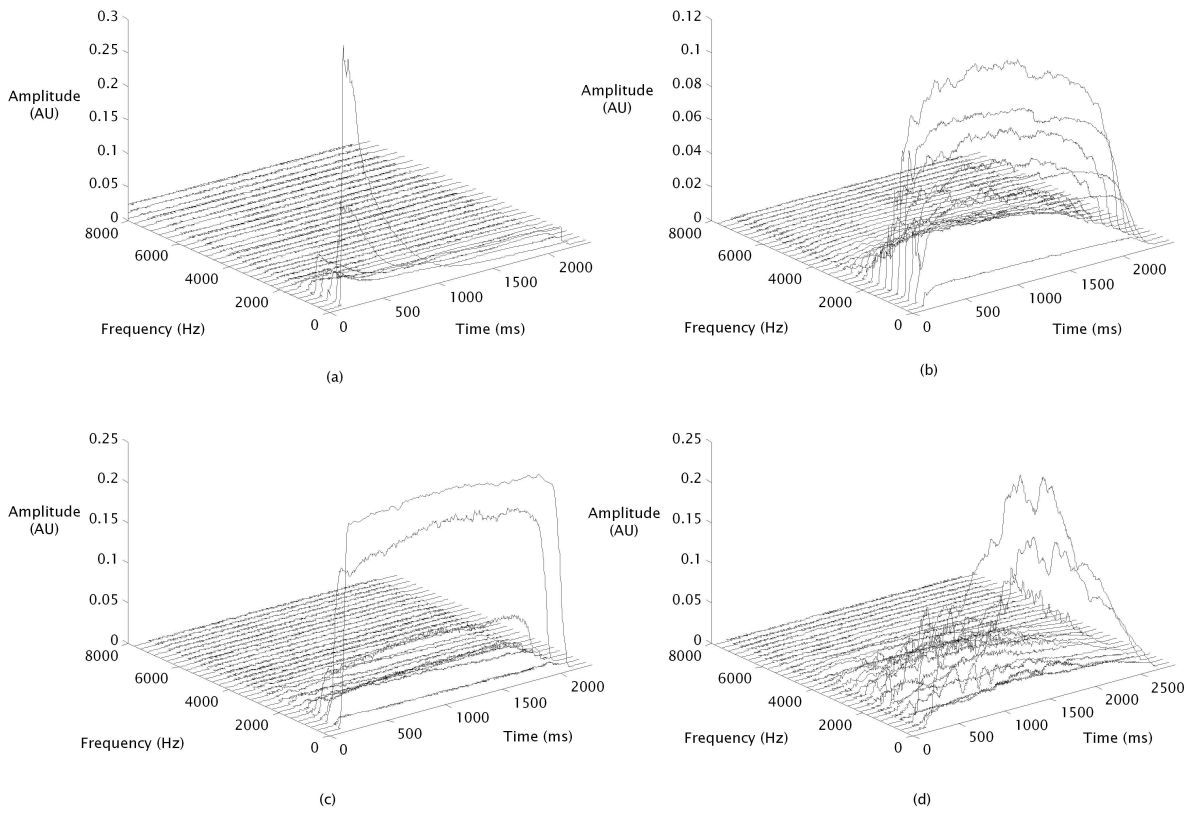

Figure 3.3.
Additive parameters for (a) the piano, (b) the trumpet, (c) the clarinet and (d) the violin.

3.3.3 Spectral envelope parameters

As discussed in section 2.3.1 in chapter 2, the spectral envelope is considered one of the most important features in defining the timbre of a musical instrument sound. Using the additive parameters as extracted in section 3.3.2, the spectral envelopes for each sound can be calculated by finding the maximum amplitude, a_k, of each partial, k. This results in the following spectral envelopes for each of the four musical instrument sounds as a function of partial index, as illustrated in figure 3.4.

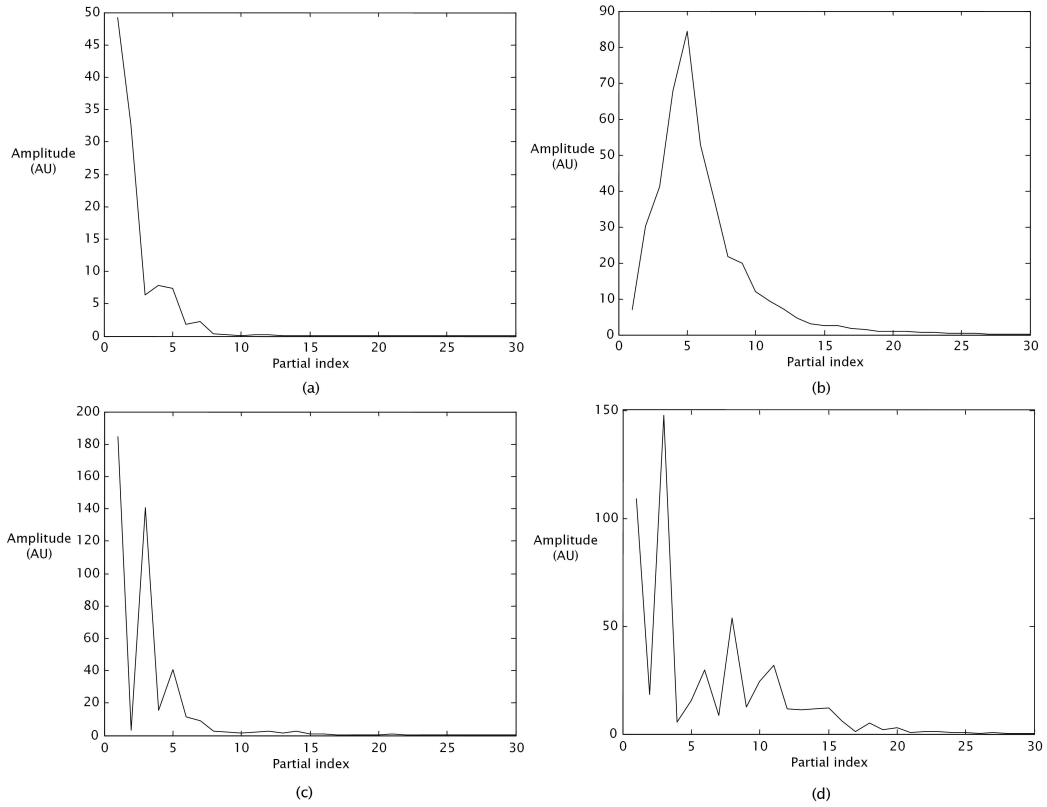

Figure 3.4.
Spectral envelopes represented as a function of the partial index for (a) the piano, (b) the trumpet, (c) the clarinet and (d) the violin, extracted from the additive parameters of figure 3.3.

Noticeable features from figure 3.4 include the slope of the envelope in each case, as well as the amplitude variations or irregularities of the spectrum. These and other important timbre features can be extracted from the spectral envelopes, based on the work of Jensen (1999a), and are discussed in sections 3.3.3.1 to 3.3.3.4.

3.3.3.1 Brightness

The brightness or spectral centroid is calculated and modelled by Jensen (1999a) from the spectral envelope from equation 3.4 as

$$brightness = \frac{\sum_{k=1}^{N} k.a_k}{\sum_{k=1}^{N} a_k}, \tag{3.4}$$

where N is the total number of partial components of the sound that are used to model the timbre. This brightness value is closely related to the attribute of sharpness, and is correlated with the subjective quality of brightness (McAdams *et al.*, 1995), which can be used to describe a sound as being "sharp" or "bright", compared to "dull". Typical brightness values, as extracted for the four musical instrument sounds of figure 3.1, are around 2.3 for the piano, 6.6 for the trumpet, 3.7 for the clarinet, and 6.5 for the violin. In the event that the partial index, k, in equation 3.4 is replaced with the frequency of the particular partial, the brightness would be expressed in Hz.

3.3.3.2 Irregularity

The irregularity of the spectrum of a musical sound has been found to be an important timbre feature (Krimphoff *et al.*, 1994; Caclin *et al.*, 2005). In the log domain, irregularity can be calculated as in equation 3.5, as the sum of the partial amplitude less the mean of the preceding, same and next partial amplitude.

$$\text{irregularity} = \sum_{k=2}^{N-1} \left| a_k - \frac{a_{k-1} + a_k + a_{k+1}}{3} \right| \tag{3.5}$$

Alternatively, irregularity can be calculated as the sum of the squared difference in amplitude

between adjacent partials, as shown in equation 3.6 by

$$\text{irregularity} = \frac{\sum_{k=1}^{N} (a_k - a_{k+1})^2}{\sum_{k=1}^{N} a_k^2}, \tag{3.6}$$

where the N+1 partial is set to zero. In general, the irregularity value calculated by equation 3.6 is below 1, and it is always below 2.

3.3.3.3 Tristimulus

The tristimulus values can be viewed as the equivalent of the colour attributes of vision (Jensen, 1999a) and can be used to investigate the transient behaviour of musical sounds. The values for tristimulus 1, 2 and 3 can be calculated as in equations 3.7, 3.8 and 3.9, respectively. The sum of the three tristimulus values equals 1.

$$\text{tristimulus 1} = \frac{a_1}{\sum_{k=1}^{N} a_k} \tag{3.7}$$

$$\text{tristimulus 2} = \frac{a_2 + a_3 + a_4}{\sum_{k=1}^{N} a_k} \tag{3.8}$$

$$\text{tristimulus 3} = \frac{\sum_{k=5}^{N} a_k}{\sum_{k=1}^{N} a_k} \tag{3.9}$$

For the purpose of illustration, a tristimulus diagram with tristimulus 2 as a function of tristimulus 3 is usually constructed. In such a diagram, the three corners are indicative of the partial strength distribution, as shown in figure 3.5 for a tristimulus diagram of the 10 musical instruments used in this study.

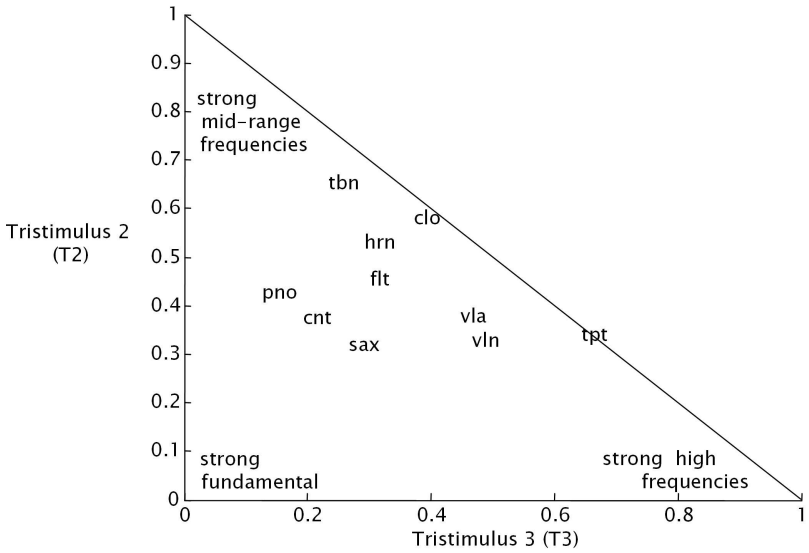

Figure 3.5.
Tristimulus values shown for 10 musical instrument sounds. The three corners denote strong fundamental partials, strong mid-range partials and strong high-frequency partials. The abbreviations for the 10 instruments are defined as piano (pno), trumpet (tpt), French horn (hrn), trombone (tbn), clarinet (cnt), saxophone (sax), flute (flt), violin (vln), cello (clo) and viola (vla).

3.3.3.4 Odd and even relationships

The odd and even relationship has been used to investigate instruments such as the clarinet, where the energy of the even partials is less than that of the odd partials (Jensen, 1999a; Gfeller *et al.*, 1998). The calculation of the odd parameter does not include the fundamental partial, so as to avoid too high a correlation between the odd parameter and the tristimulus 1 parameter. Equations 3.10 and 3.11 show the calculations for the odd and even relationships respectively.

$$\text{odd} = \frac{\sum_{k=2}^{N/2} a_{2k-1}}{\sum_{k=1}^{N} a_k} \tag{3.10}$$

$$\text{even} = \frac{\sum_{k=1}^{N/2} a_{2k}}{\sum_{k=1}^{N} a_k} \tag{3.11}$$

The sum of tristimulus 1, odd and even, equals 1; thus, only the odd parameter needs to be calculated and saved when modelling timbre with as few parameters as possible.

3.3.4 Amplitude envelope times: attack, sustain and release

The amplitude envelope is another important attribute of timbre, as discussed in section 2.3.1, and defines the evolution of the amplitude of a sound over time. To model the amplitude envelope of a musical sound, a number of steps need to be implemented, as discussed in detail by Jensen (1999a). The procedure is briefly described in the following paragraphs.

As a first step, a Gaussian window is convolved with each partial of the instrument sound to be modelled to obtain a smoothed version of each partial. From the smoothed partial, start of attack (soa), end of attack (eoa), start of release (sor) and end of release (eor) points of the partial envelope can be estimated. This is achieved by finding the maximum and minimum points of the first derivative version of the smoothed partial. The maximum value corresponds to the middle of the attack segment, while the minimum value corresponds to the middle of the release segment. From these middle points, the split points of the smoothed partial, soa, eoa, sor and eor, can be calculated as a percentage of the middle point on either side (usually set at 10 % above or below the middle point).

The zero-crossings of the third derivative of the smoothed partial correspond to the start and end points of the segments (Lindeberg, 1996). The zero-crossing points closest to the attack and release values found from the first derivative values are then used as the initial smoothed soa, eoa, sor and eor values. However, the split point times found from the smoothed version of the partial obviously do not correspond to the original slope times, so these must be traced back to correspond to the original, unsmoothed partial. This is done by following the split points from the smoothed version to the unsmoothed version of the partial in steps of different degrees of smoothing of the partial. Steps of smoothing are implemented by Gaussian windows with changing α values, where a small α value corresponds to a smooth

signal and a high α value corresponds to an unsmoothed signal. Thus, for each smoothing step, the zero-crossing values closest to the previous (more smooth) zero-crossing values are followed through to the unsmoothed case, giving the correct slope times.

The process described above was implemented as a preliminary approach to this study. However, although the calculation of these split times is important in modelling the timbre of musical instrument sounds, these were not included in this study. Rather, a logarithm of the rise time of the sound envelope was found to be an important feature for timbre perception and was used instead. This will be discussed in detail in section 3.4.2.

In the modelling of timbre, noise components or irregularities are often added to the envelopes, in terms of shimmer (irregularities of the amplitudes of the partials) and jitter (irregularities on the frequencies of the partials), as discussed in section 2.3.3. Again, in this study this approach was not followed, as only three primary timbre features were focussed on, with noise components not playing as important a role.

3.3.5 Resynthesis: summation of the sinusoids

The additive analysis, as explained in section 2.3.1 in chapter 2, involves the association of a number of sinusoids with a sound. The time varying amplitudes, $a_k(t)$ and frequencies $f_k(t)$ of the N partials of the sound are estimated, from which the original sound can be resynthesised with a high degree of realism in terms of sinusoids. This is achieved by implementing equation 3.12 (Jensen, 2002b; Jensen, 2001; Andersen and Jensen, 2001) as

$$s(t) = \sum_{k=1}^{N} a_k(t).sin(\phi_k(t)), \tag{3.12}$$

where a summation of the N sinusoids is performed over all time, *t*, to produce the resynthesised sound, *s(t)*, in time. The integral of the frequency in equation 3.12 is the phase, $\phi_k(t)$ of the particular sinusoid, defined by equation 3.13 as

$$\phi_k(t) = 2\pi \int_{\tau=0}^{t} f_k(\tau) d\tau. \tag{3.13}$$

For the correct implementation of the resynthesis, the integral of equation 3.13 was approximated using a summation of the frequencies up until τ. The midpoint rule was used as the summation method to approximate the integral, due to the low errors incurred by this method (Stewart, 1999).

3.4 IMPORTANT TIMBRE FEATURES

Many studies on timbre perception features for acoustic hearing have been performed, as discussed in 2.3.1. A number of possible acoustic correlates of timbre-space dimensions have been presented in the psychoacoustic literature, including the spectral centre of gravity or spectral centroid, various forms of the attack time, the spectral flux, and the spectral fine structure of the sound (Caclin *et al.*, 2005). An important conclusion that can be drawn from literature is that three main important features for the perception of timbre in acoustic hearing can be assumed . For the purpose of this study, the features for acoustic hearing will be assumed to be the important features for hearing in CI listeners. The three most important features, as investigated and summarised by Krimphoff *et al.* (1994), McAdams *et al.* (1995) and Caclin *et al.* (2005) that were thus extracted from the acoustically modelled sounds are:

- the spectral centroid or brightness (B)
- the logarithm of the rise time (LRT)
- the spectral irregularity (IRR)

The calculations implemented to extract each of these features are discussed in sections 3.4.1 to 3.4.3 that follow. Although the units of the three important timbre features will be defined in the calculations that follow, it should be noted that throughout the remainder of the dissertation these features will be referred to without units for the ease of illustration, as well as for the sake of consistency with existing literature involving these parameters.

3.4.1 Brightness or spectral centroid

The spectral centroid calculations are implemented according to the methods described by Krimphoff *et al.* (1994), McAdams *et al.* (1995), Lakatos (2000), Iverson and Krumhansl (1993) and Beauchamp and Lakatos (2002). These calculations find the average spectral centroid over the duration of the tone, using instantaneous spectral centroid values calculated over individual time windows. This gives the spectral centroid as a function of time, B_t, from which a time-average can be found as B. Beauchamp (1993) implements an algorithm to calculate this feature. The spectral centroid values B_t and B can be expressed by equations 3.14 and 3.15 as

$$B_t = \frac{\sum_{k=1}^{N} k.a_k(t)}{\sum_{k=1}^{N} a_k(t)} \tag{3.14}$$

and

$$\mathrm{B} = \frac{\sum_{t=1}^{T} B_t}{T}, \tag{3.15}$$

where k is the partial index and $a_k(t)$ is the amplitude of each partial for each time window, t, with T being the total number of time windows. The time windows are represented by each element of the matrix that holds the partial amplitude values. The above equations show that the units of B can be defined as the partial number or index. The global spectrum with the B values indicated for each of the four instruments of figure 3.1 is illustrated in figure 3.6. Only the first 15 partial components are shown, for the purpose of illustration.

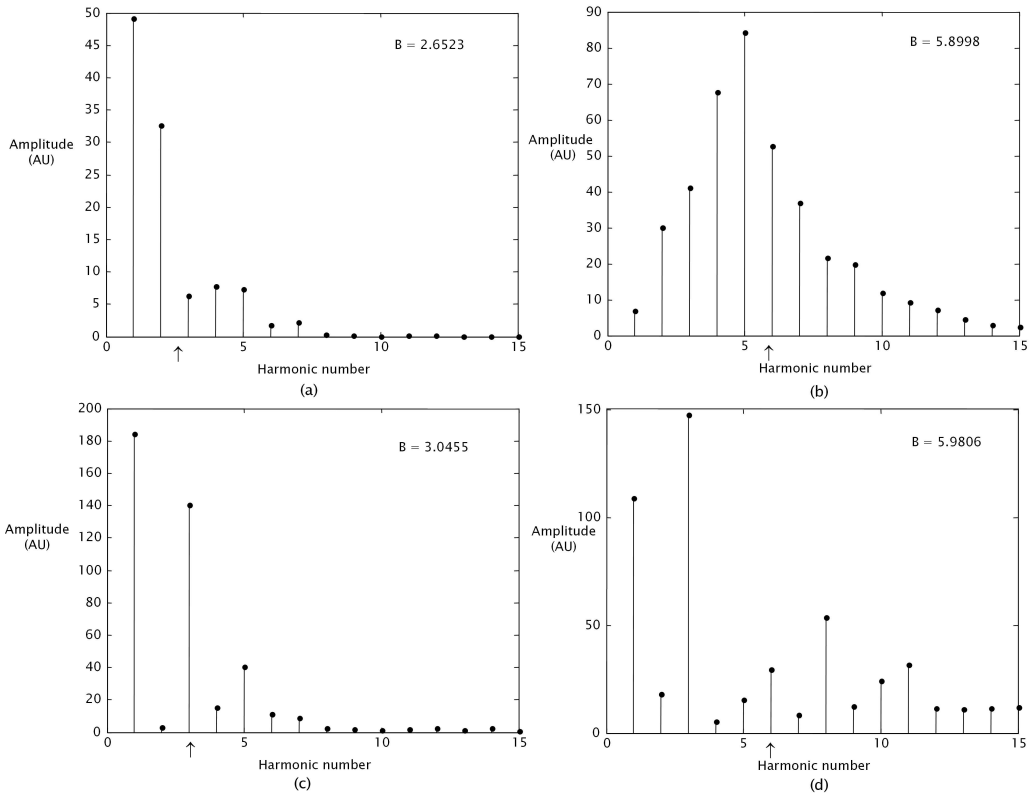

Figure 3.6.
B values indicated for each of the four musical instrument sounds: (a) piano, (b) trumpet, (c) clarinet and (d) violin. The arrows indicate the position of the spectral centroid in relation to the global spectrum of each instrument, with the units of B as defined in section 3.4.1.

3.4.2 Logarithm of rise time

As discussed in section 2.3.1, the rise time of an instrument sound is an important feature for timbre perception, as it distinguishes impulsive tones from sustained tones (Caclin *et al.*, 2005). Krimphoff *et al.* (1994) and McAdams *et al.* (1995) conclude that the LRT value, the logarithm of the time taken for the sound to reach a maximum from the time it reaches 10 % of the maximum, correctly defines this timbre dimension. The envelope of the sound signal from which LRT can be calculated is obtained by finding a quadratic sum of all the partial amplitudes over the duration of the sound and finding the square root of this sum. Alternatively, as discussed by Krimphoff *et al.* (1994), a linear sum of the partials can be used.

The calculation for the temporal envelope, $Env(t)$, of the sound is found by equation 3.16 as

$$Env(t) = \sqrt{\sum_{k=1}^{N} (a_k(t)^2)}. \tag{3.16}$$

LRT can be found from the sound signal envelope, $Env(t)$ and can be calculated from equation 3.17 as

$$\text{LRT} = \log(t_{max} - t_{0.1max}), \tag{3.17}$$

where $t_{0.1max}$ and t_{max} are the times (in seconds) at which the temporal envelope of the sound, $Env(t)$, reaches 10 % of its maximum value and its maximum value, respectively. The units of LRT are thus given as the logarithm of time in seconds, or log(s). Figure 3.7 shows the LRT values calculated for each of the four instruments of figure 3.1, as indicated on the sound envelopes.

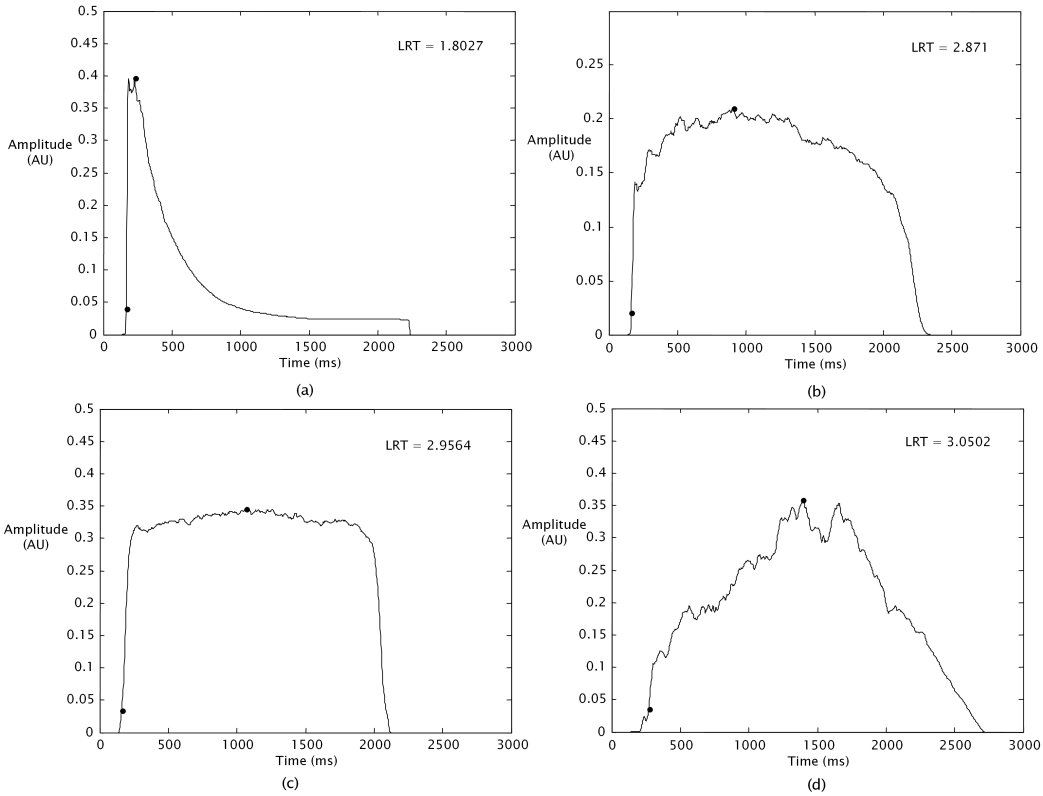

Figure 3.7.
LRT values indicated for each of the four musical instrument sounds: (a) piano, (b) trumpet, (c) clarinet and (d) violin. The amplitude envelopes of each of the sounds are shown. The two filled circles indicate the start of the rise time and end of the rise time in each case, with the units of LRT as defined in section 3.4.2.

3.4.3 Spectral irregularity

The third most important feature in timbre perception is the irregularity in the spectrum of the sound, as discussed by Krimphoff *et al.* (1994), Caclin *et al.* (2005) and Beauchamp and Lakatos (2002). Caclin *et al.* (2005) discuss how this feature involves the attenuation of even harmonics relative to odd harmonics. Mathematically, IRR can be defined as the SD of a running mean of three adjacent partial amplitudes from a global spectral envelope; that is, the spectral envelope over the entire duration of the sound (McAdams *et al.*, 1995). The logarithm of this value then gives the IRR value, as expressed by Krimphoff *et al.* (1994) and shown by equation 3.18 as

$$\mathrm{IRR} = \log\left(\sum_{k=2}^{N-1}\left|20\log(a_k) - \frac{20\log(a_{k+1}) + 20\log(a_k) + 20\log(a_{k-1})}{3}\right|\right), \tag{3.18}$$

where a_k is the sum of the amplitudes for partial k over all time. This equation indicates that the units of IRR are defined as the logarithm of decibels, or log(dB).The IRR values calculated for each of the four musical instrument sounds of figure 3.1 are shown in figure 3.8, with the corresponding logarithm of the global spectrum for each sound, from which IRR is calculated. The first 20 partial components are shown for each spectrum for the purpose of illustration.

It can be noted that the spectrum of the trumpet is very smooth, with little to no irregularity in adjacent harmonics, as the components follow similar patterns. This indicates a low IRR value in contrast with the clarinet, for example, where the spectrum is jagged and the odd and even harmonics differ substantially, thus giving a higher IRR value.

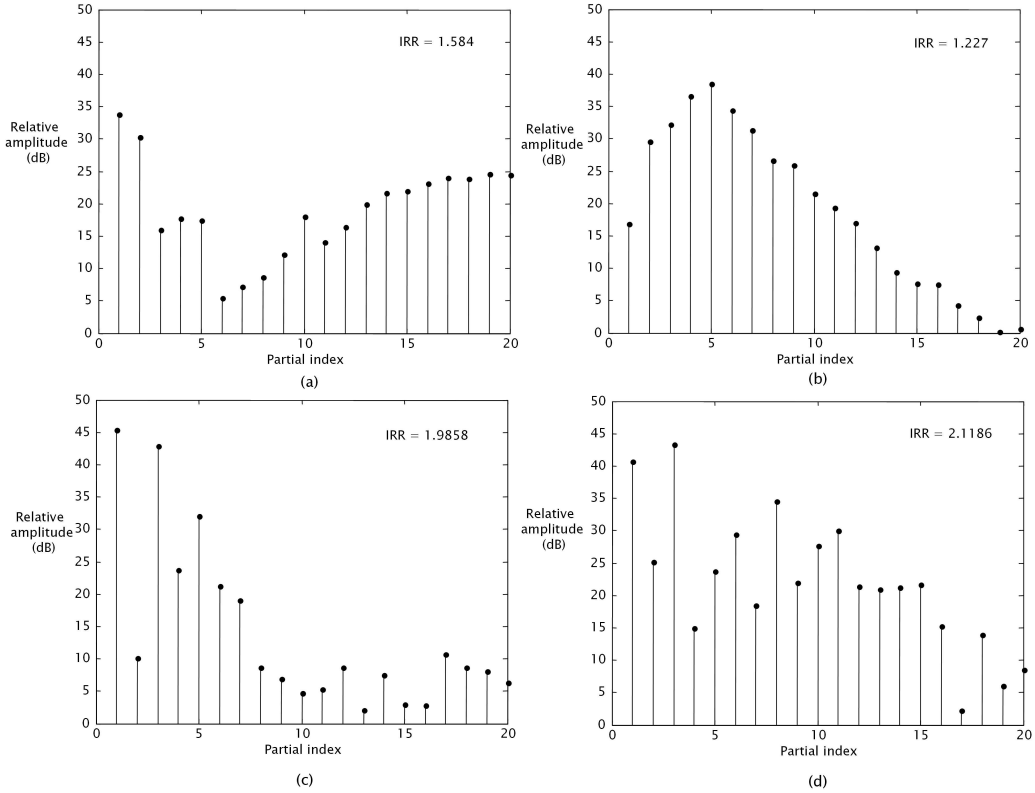

Figure 3.8.
IRR values indicated for each of the four musical instrument sounds: (a) piano, (b) trumpet, (c) clarinet and (d) violin. The relative logarithm amplitudes of the sound spectra are shown, from which the IRR values are calculated and given with units as defined in section 3.4.3.

3.5 DEVELOPMENT OF THE ACOUSTIC MODEL

The approach followed in developing the acoustic model was to separate the model into the signal processing aspects and the biophysical characteristics of the electrode-neural interface. The biophysical characteristics of the electrode-neural interface are more complicated to implement generically, and many assumptions must be made in order to model this part of the acoustic simulation. For the acoustic model implemented in this study, emphasis was placed on the processing part of the model, as the correct implementation of the processor functioning is a necessity in understanding how the sound signal is affected. To obtain an accurate simulation of the CI speech processor, the Nucleus Matlab Toolbox (NMT) from Cochlear Pty Ltd [2] was used.

The Matlab toolbox developed by Cochlear Pty Ltd was designed to emulate the processing of speech by a CI. The toolbox allows for the generation of current signals that can be applied directly as the stimulus to a CI electrode array, facilitating experiments performed with CI users. By examining the processing steps of the NMT, the important steps to be implemented for the acoustic simulation could be extracted.

In the sections that follow, the development of the acoustic model is described in detail, based on the processing steps presented in the NMT as well as on previously developed models.

[2] www.cochlear.com

3.5.1 Processing steps of the Nucleus speech processor

The Nucleus CI processor incorporates different types of speech processing strategies, which are implemented by the NMT. The CIS strategy focusses on the temporal information of the sound signal, while the SPEAK and ACE strategies focus on the spectral information. The ACE strategy was selected as the approach to follow in the implementation of the acoustic model. In this strategy, the incoming sound is usually divided into 22 frequency bands or channels, and the six channels with the highest energy content for a given time window are used for stimulation during that time.

In the ACE processing strategy, the sound signal is divided into fixed 8 ms time windows with a 75 % overlap. These time windows are weighted by a Hanning window to avoid abrupt transitions in the time domain of the sound signal, and thus reduce the resulting spectral spread of the sound spectrum. Following this, the signal is divided into frequency bands using a FFT, whereby the frequency bins for each of the strategies are predetermined using a filter analysis table (FAT). Alternatively, a number of bandpass filters may be used instead of frequency bins to filter the sound into channels. This method has been used in existing acoustic models, as discussed by Loizou (1998), and is the approach followed in this model.

Once the sound has been filtered into channels for each time window, the energy content of each band for a specific time window is determined. The length of one time window is 128 samples, corresponding to 8 ms for a sampling rate of 16 kHz. This is the standard sampling rate of the analogue-to-digital converter (ADC) of the processor that digitises the analogue input sound. As a result of the 75 % overlap of the time windows, new samples will be available every 2 ms. For each of these time windows, the envelopes of the filtered signals are determined, representing an estimate of the instantaneous power in the corresponding channel (Cochlear Pty Ltd, 2002).

In the ACE strategy, only the subset of channels with the highest energy content for a specific time window are selected, and the corresponding channels are stimulated sequentially for that time. The maximum overall stimulation rate of the Nucleus speech processor is 14400 pps. The stimulation rate of an individual electrode is dependent on the number of channels in use. For example, if six channels are selected for each stimulation cycle, the resulting maximum stimulation rate of a single electrode would be 2400 pps (14400 pps divided by six channels). A typical setting for the ACE strategy is to select eight maxima out of the 22 channels and stimulate at a rate of 1200 Hz (Cochlear Pty Ltd, 2002).

Once the subset of channel maxima has been selected, the calculated energy levels for these channels are mapped to current levels which will be used to stimulate the nerve cells in the cochlea through an electrode array. These current levels must adhere to a range between a minimum current level, known as the threshold level, T, and a maximum current level, known as the comfort level, C. The threshold level is the minimum current value that produces a stimulus that is only just audible. The comfort level is the maximum current value that can be used just before the stimulus becomes uncomfortably loud. Current levels that fall outside of the range are clipped in the NMT to ensure that all the values fall within the C and T levels. The C and T levels are user dependent for each electrode pair in a CI and can be changed in the NMT according to individual requirements. The NMT implements a logarithmic function, which is referred to as a loudness growth function (LGF), that maps the energy levels to current values between the C and T levels.

Once the current levels have been obtained, the selected channels are stimulated, with default activation in the NMT starting from the most basal position and moving to the most apical position. This approach is also followed for the developed acoustic model, by ordering the channels according to their centre frequencies. Finally, the current values are mapped to electrodes along the array that will stimulate specific places along the cochlea.

3.5.2 Processing steps implemented in the acoustic model

The processing steps in the acoustic model must be as true to the actual processing performed in a CI as possible. All the processing steps were implemented in Matlab, following the processing methods of the NMT as described in the previous section. The instrument sound signals used for this study were processed by Matlab code, with the output saved as a .wav file for each sound.

The block diagram in figure 3.9 illustrates the steps of the acoustic model, clearly showing the processing aspects (labelled as Processor model) and the biophysical characteristics of the electrode-neural interface (labelled as Biophysical model) of the acoustic model. A description of each functional block is given in the sections that follow, with the processing aspects continuing in section 3.5.2 and a discussion of the biophysical characteristics of the electrode-neural interface following in section 3.5.3. The shaded blocks in figure 3.9 indicate the biophysical characteristics of the electrode-neural interface that were not included in the final implementation of the acoustic model in this study, as discussed in section 3.5.3.

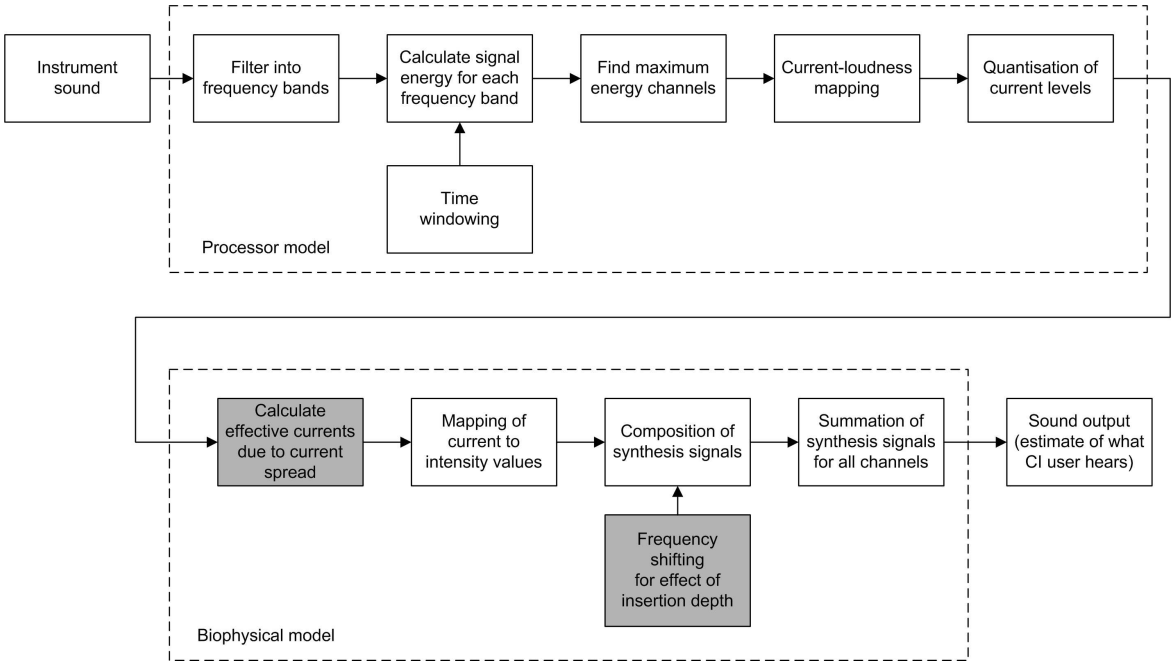

Figure 3.9.
Block diagram illustrating the implementation of the acoustic model. The biophysical model characteristics of the electrode-neural interface indicated by the shaded blocks were not included in the final implementation.

Illustrations of each processing step are given where possible for a single 2 s long C4 note of a piano sound. The original instrument sound signal is shown in the time and frequency domain in figures 3.1(a) and 3.2(a), respectively, which are reproduced here in figures 3.10 and 3.11 for the purpose of comparison with the processing steps illustrated in this section.

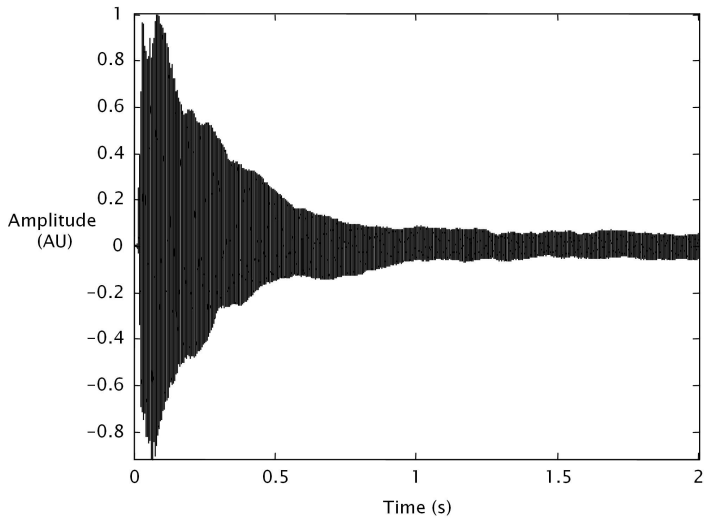

Figure 3.10.
Time domain representation of the unprocessed piano sound as shown in figure 3.1(a) previously.

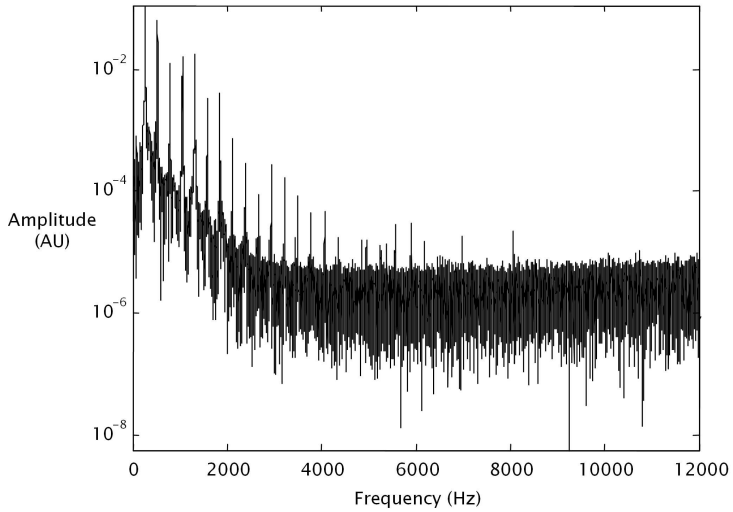

Figure 3.11.
Frequency domain representation of the unprocessed piano sound as shown in figure 3.2(a) previously, displayed up to 12 kHz.

Signal pre-emphasis generally forms part of CI processing and CI simulations as the first processing step. This process de-emphasises the low-frequency content of the sound, so that peak-picking strategies such as ACE are less low-pass in nature. This promotes higher frequency channels to be selected and included in the stimulation pattern. It was decided not to include signal pre-emphasis in the implementation of the acoustic model in this study, as this is usually implemented with speech signals in mind. Speech contains important higher frequency cues and for these not to be lost, pre-emphasis is employed. However, as this study involved musical instrument sounds, with the lower frequency elements being the most prominent, pre-emphasis of the sound signals was omitted. Additionally, the focus of this study was not on the acoustic model implementation, as will be discussed in a later stage, and thus processing phases such as signal pre-emphasis and biophysical characteristics of the electrode-neural interface were omitted from the acoustic model.

3.5.2.1 Bank of bandpass filters

The first step of the processing side of the acoustic model is to filter the original musical instrument sounds, which are read from a .wav file, into 22 frequency bands. The filter configurations that will be used in this study are the standard bandpass filters used by the ACE strategy. Laneau and Wouters (2004) investigate different filter bank configurations employed in CIs, including the configuration usually implemented for the ACE strategy, and show how they affect fundamental frequency discrimination.

The filter allocation tables for the ACE strategy, also corresponding to to the ACE implementation example given by the NMT, were implemented. Table 3.1 shows the values for the lower and upper cut-off frequencies as well as the centre frequencies of the bandpass filters, as calculated for the ACE strategy for 22 frequency bands.

The 22 bandpass filters were chosen to be sixth-order Butterworth filters, as a result of the flat bandpass response obtained by this type of filter. The number of filters corresponds to the number of possible places in the cochlea that may be stimulated with the electrode array, with a 22 electrode array commonly being found in Nucleus CIs.

Table 3.1.
-3 dB cut-off and centre frequencies for the bandpass filters.

Channel	Lower cut-off frequency (Hz)	Upper cut-off frequency (Hz)	Centre frequency (Hz)
1	188	313	250
2	313	438	375
3	438	563	500
4	563	688	625
5	688	813	750
6	813	938	875
7	938	1063	1000
8	1063	1188	1125
9	1188	1313	1250
10	1313	1563	1437.5
11	1563	1813	1687.5
12	1813	2063	1937.5
13	2063	2313	2187.5
14	2313	2688	2500
15	2688	3063	2875
16	3063	3563	3312.5
17	3563	4063	3812.5
18	4063	4688	4375
19	4688	5313	5000
20	5313	6063	5687.5
21	6063	6938	6500
22	6938	7938	7437.5

An illustration of the frequency response of the filter bank configuration is shown in figure 3.12. The filters were implemented as infinite impulse response (IIR) filters. Figure 3.13 gives an example of the transfer function of the bandpass filter implemented for channel 3, with the resulting instrument sound signal as filtered by this channel shown in the time domain (figure 3.14) and the frequency domain (figure 3.15).

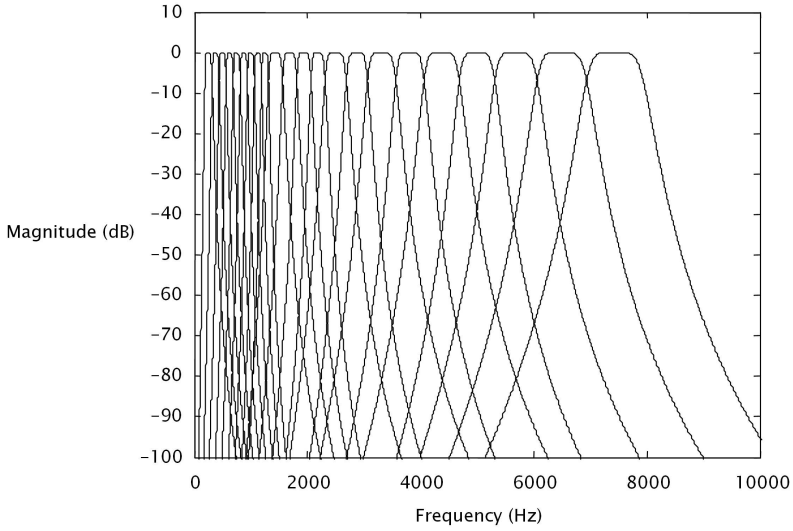

Figure 3.12.
Frequency response of the filter bank configuration that will be implemented in the acoustic model.

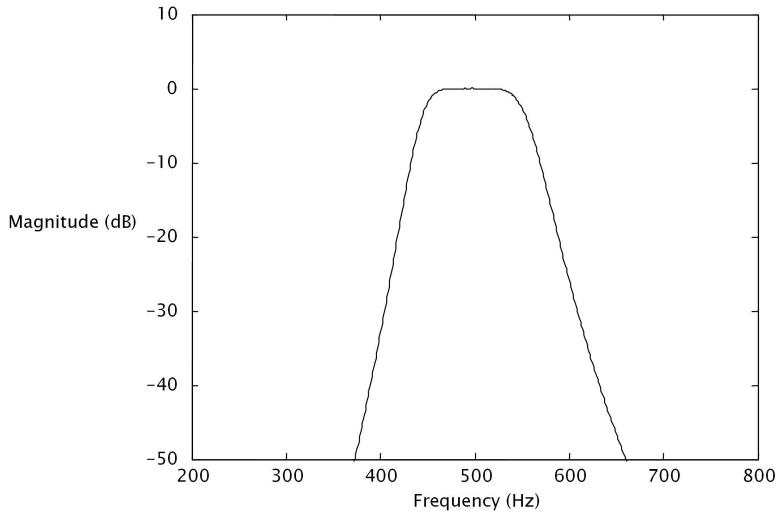

Figure 3.13.
Transfer function of the bandpass filter for channel 3.

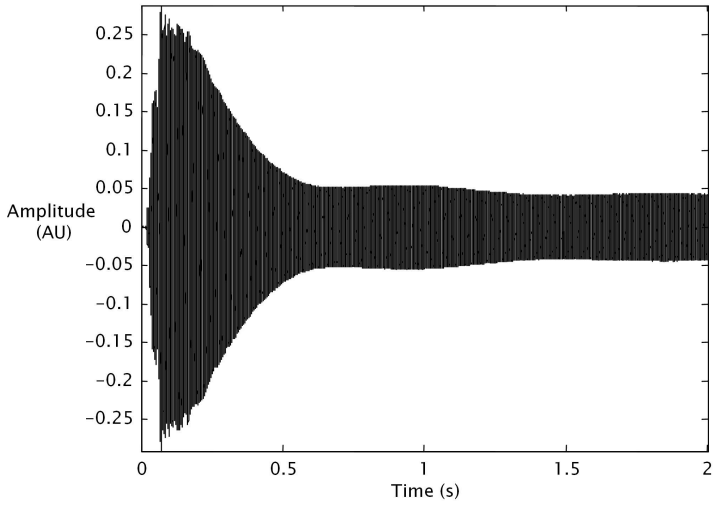

Figure 3.14.
Time domain representation of the bandpass filtered piano sound through channel 3.

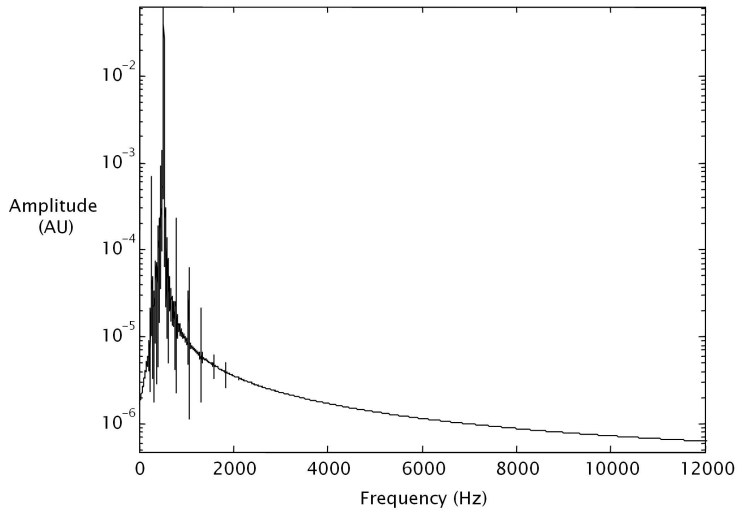

Figure 3.15.
Frequency domain representation of the bandpass filtered piano sound through channel 3.

3.5.2.2 Energy calculations in each channel

Once the musical instrument sound has been filtered into 22 channels, a representation of the energy in each channel is calculated. This is achieved by extracting the envelope of each channel by means of full wave rectification and lowpass filtering, followed by root-mean-square (RMS) calculations. The envelopes of each of the 22 bands can be obtained by first implementing full wave rectification of the signals of each channel, calculated by equation 3.19 as

$$A_{\text{channel FWR}} = |A_{\text{channel}}|\,, \tag{3.19}$$

where A_{channel} is the signal amplitude of a specific channel or band and $A_{\text{channel FWR}}$ is the resulting full wave rectified signal amplitude for the channel, as illustrated in figure 3.16.

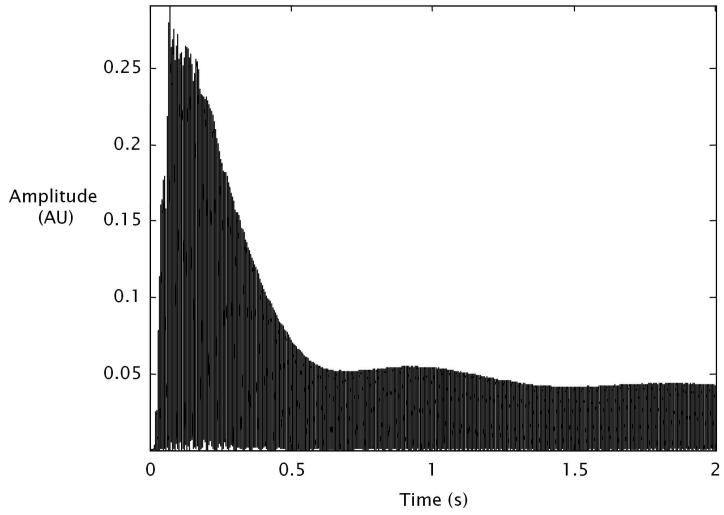

Figure 3.16.
Example of time domain representation of the full wave rectified piano sound for channel 3.

Full wave rectification causes additional frequency components of the signal to appear at 0 Hz and at double frequencies of the signal. By lowpass filtering the rectified signal data, only the lower frequency components will remain. This results in an envelope of the rectified signal, from which a representation of the energy content of the signal can be determined by RMS calculations.

A second order Butterworth lowpass filter with a -3 dB cut-off frequency of 125 Hz was used to filter the rectified signal of each channel. This ensures that the double frequency components generated as a result of signal rectification are removed. An illustration of the transfer function of the lowpass filter that was implemented as an IIR filter in Matlab is shown in figure 3.17, followed by an illustration of the resulting envelope of the lowpass filtered signal through channel 3 in figure 3.18.

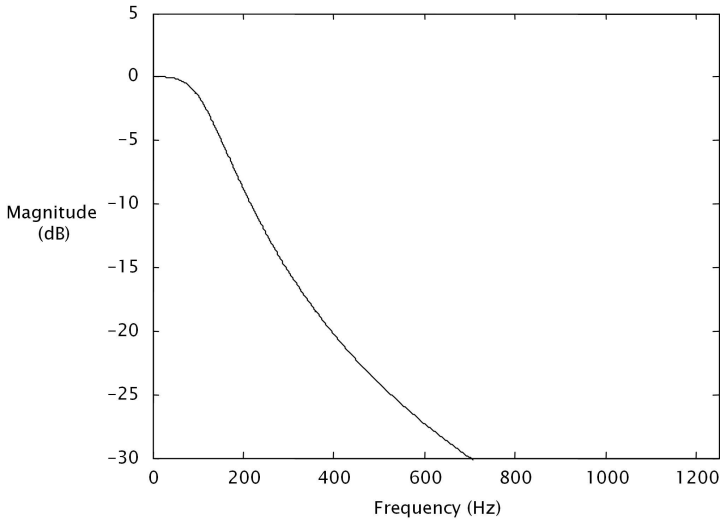

Figure 3.17.
Transfer function of 125 Hz lowpass filter used to filter the full wave rectified signal of each channel.

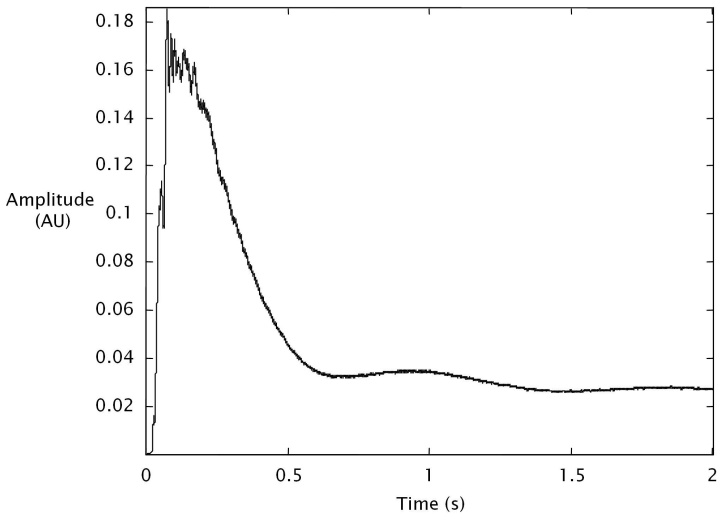

Figure 3.18.
Example of time domain representation of full wave rectified and lowpass filtered sound for channel 3.

3.5.2.3 Root mean square calculations

The energy content of the signal can be represented by the calculated RMS of the rectified and lowpass filtered data in each channel. Each channel is divided into a number of time windows with an overlap of 75 %, and the RMS is calculated for each of these time windows.

To compensate for the spectral spread that is introduced by dividing the signal into windows of time, a Hanning window is used to smooth each time window of the signal before the RMS calculations are performed. This is to ensure that no abrupt signal transitions occur, and that the high frequency components are removed. The time windows are fixed to be 8 ms long, regardless of the number of channels being used or the rate of stimulation. The number of samples in an 8 ms time window depends on the sampling frequency of the original signal, and can be calculated as shown in equation 3.20 by

$$N = 8(\text{ms}) \times f_s(\text{samples/ms}), \tag{3.20}$$

where N is the number of samples in the Hanning window and f_s is the sampling frequency in kHz. The instrument sounds used in this study all have a sampling frequency of 44.1 kHz, resulting in 353 samples for each time window. The Hanning window weights for the calculated sample length are illustrated in figure 3.19.

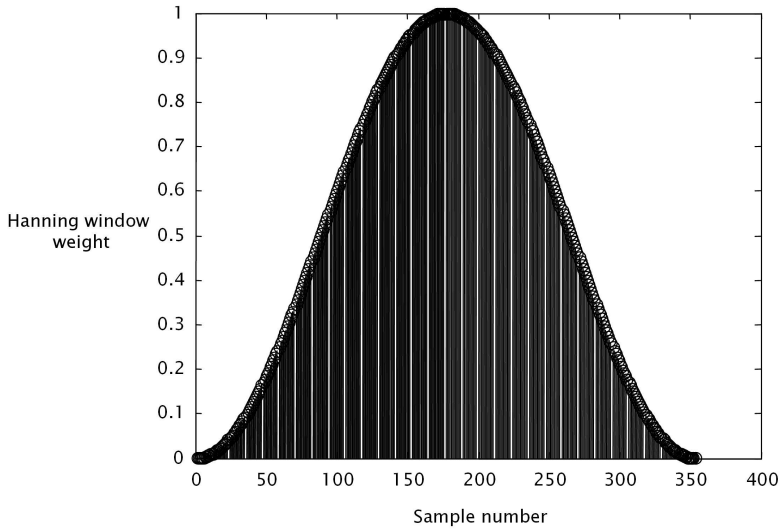

Figure 3.19.
Hanning window values for a window length of 353 samples (8 ms for a sampling frequency of 44.1 kHz).

The signal envelopes of each channel are divided into overlapping time windows of 353 samples and are multiplied by the Hanning window values of figure 3.19 for each window. The weighted time window signal envelopes can then be used to calculate the RMS values for each time window. The RMS is an indication of the energy content in a specific frequency channel for a given window of time, and can be calculated from equation 3.21 as

$$A_{\text{RMS}} = \sqrt{\frac{1}{N}\sum A_n^2} \qquad n = 1, ..., N, \tag{3.21}$$

where N is the number of samples in the time window and A_n is the amplitude of the n^{th} sample of the signal in the time window for a specific channel.

To illustrate the above-mentioned procedure, figure 3.20 shows the signal envelope of channel 3, with the resulting RMS values extracted from the signal indicated in red. Figures 3.21 and 3.22 display the calculated RMS values over 2 ms time windows.

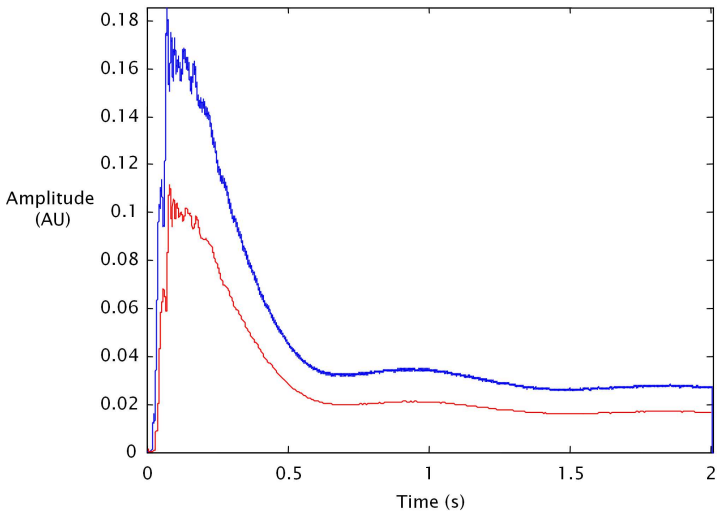

Figure 3.20.
Example of time domain representation of full wave rectified and lowpass filtered sound for channel 3, with calculated RMS values in a 2 ms window indicated in red.

3.5.2.4 Maximum energy calculations

Once the RMS values for each 2 ms time window have been calculated for each channel, the 6 channels with the highest RMS values are found for each time window. For each channel, the RMS value for a specific time window represents the energy content of that channel for that specific time window. Thus the 6 channels with the highest RMS values for a specific time window are chosen to be the stimulating channels for that particular time window. In this way, 6 channels and their corresponding RMS values are selected for each time window across the duration of the sound. The 6 maximum RMS values found for each time window can then be mapped to current amplitudes that will be used as stimuli for each time window.

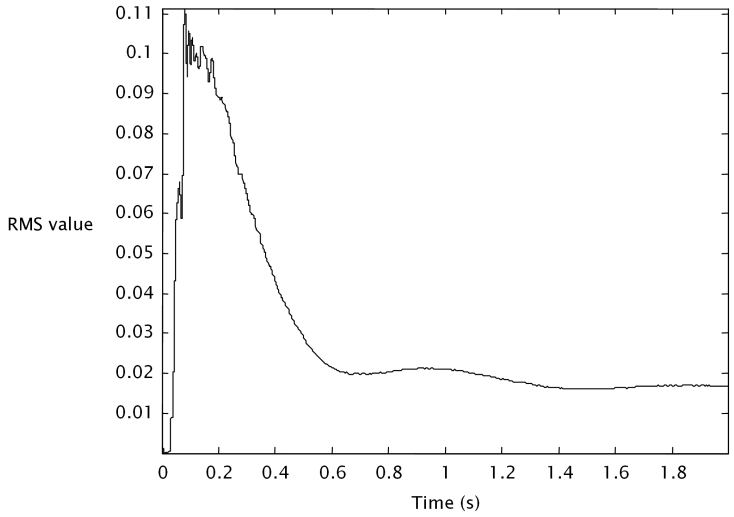

Figure 3.21.
Calculated RMS values for sound signal of channel 3, as shown in figure 3.18.

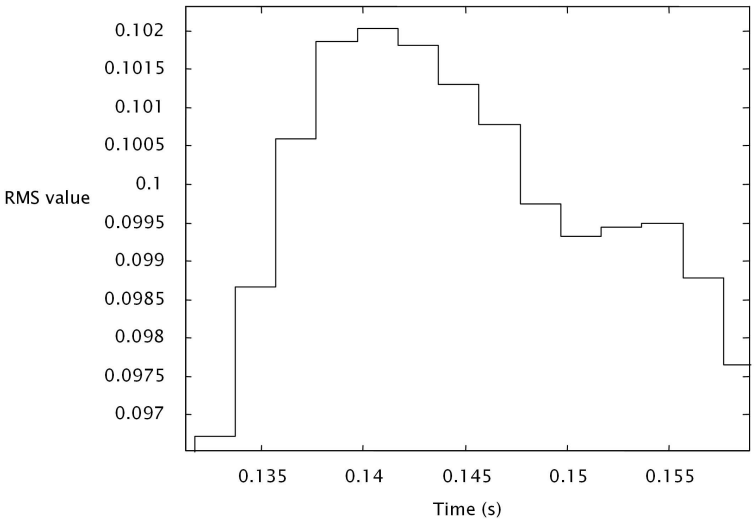

Figure 3.22.
Close-up of RMS values from figure 3.21, illustrating that the RMS values remain constant for 2 ms, which is the effective window length due to the overlap of the time windows.

3.5.2.5 Current to loudness mapping

The calculated RMS values that represent the energy content in each channel for each window of time must be translated into current magnitudes to be used as stimuli. The current values are used in the remaining steps of the model, including the reconstruction of the sound signal using sinusoidal signals.

The RMS values obtained must first be scaled to the input dynamic range (IDR). The input dynamic range for CI listeners is approximately 30 dB or less for an optimal microphone input (Van Hoesel and Tyler, 2003; Zeng *et al.*, 2002; Shannon, 1983). Assuming a generalised dynamic range of 30 dB as set for the CI listener, this can be interpreted as base and saturation input levels of 4 and 150, respectively. The maximum RMS value obtained over all channels is set to the maximum input magnitude of 150. The remaining RMS values are scaled accordingly. In this way, no values will exceed the saturation level, and clipping of input values to the comfort current level, C, will be prevented.

The input values ranging between the base and saturation levels are mapped to current levels by means of a logarithmic function shown in equation 3.22 by

$$I_{\text{mag}} = \frac{\log\left(1+\alpha\left(\frac{m-b}{s-b}\right)\right)}{\log(1+\alpha)}, \tag{3.22}$$

where I_{mag} is the mapped current magnitude, calculated using the RMS values scaled to the IDR magnitudes, m, the base and saturation levels, b and s, and the α parameter that controls the steepness of the curve. The parameter α is related to Q, which is known as the steepness factor and is defined as the percentage decrease in the output for a 10 dB decrease in the input. For this study, Q was set to a typical value of 20, implying a 20 % decrease in output level for a 10 dB decrease in the input. This results in a value of 416.2063 for α.

The minimum and maximum current levels, T and C, are assumed to cover a current range of 12 dB, falling within the typical dynamic range of 5 - 30 dB found for CI listeners by Shannon (1983). Thus, the maximum comfort level, C, is set as 1 mA, a typical comfort level value (Clark, 2003; Bruce, White, Irlicht, O'Leary, Dynes, Javel and Clark, 1999).

The minimum or threshold current value would then fall 12 dB below this maximum, to give a corresponding range of current values from 0.215 mA to 1 mA. If the RMS input magnitude is less than the minimum input base level of 4, the mapped current value will be clipped to the T value to ensure that all current levels fall within a 12 dB range.

An illustration of the loudness growth function implemented is given in figure 3.23, followed by an example of the resulting mapped current values for the sound signal of channel 3 in figure 3.24.

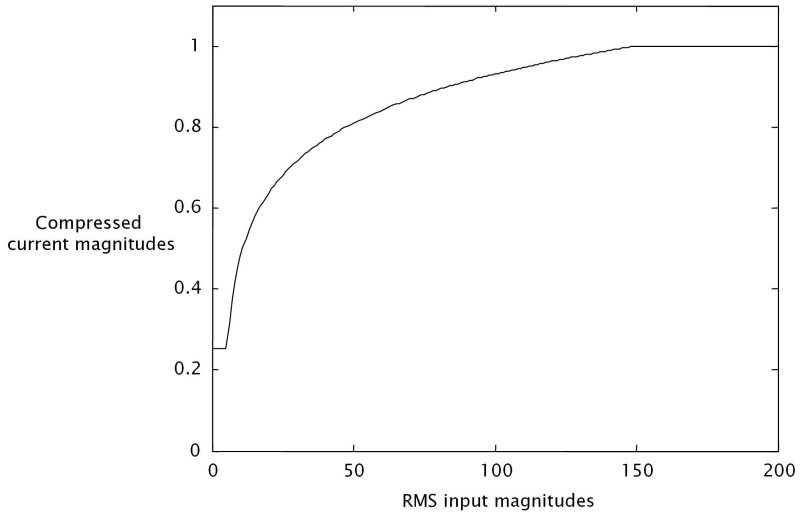

Figure 3.23.
Loudness growth function applied to the scaled RMS values to linearise the relationship between stimulus current and perceived loudness for CI listeners.

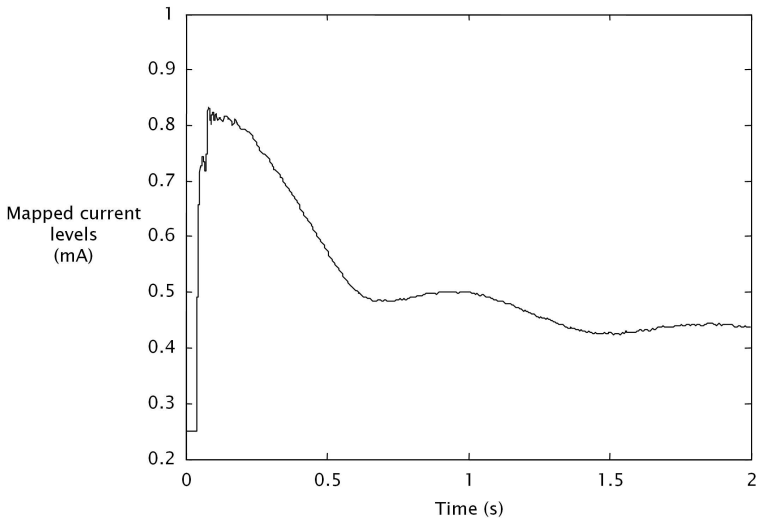

Figure 3.24.
RMS magnitudes mapped to current levels for the piano sound processed through channel 3, as in figure 3.14.

3.5.2.6 Quantisation of current levels

The quantisation of the current levels must be performed for an accurate representation of the current values that can be output by the processor. There are 236 current levels available for the Nucleus CI, and these levels span the range of current values that the current source of the implant can produce. This range is typically between 10 μ A and 1.7 mA (Clark, 2003). Using the formula given in equation 3.23 below, 256 current levels, *CL*, are converted to current values, I_{quant}, in μA, for each level.

$$I_{\text{quant}} = e^{(0.02025 \cdot CL + 2.30259)} \qquad CL = 1, 2, \ldots, 256 \tag{3.23}$$

Since only 236 levels of current can be used in the Nucleus processor, the lower current levels, corresponding to *CL* in the range of 1 to 20, may be excluded, with *CL* then ranging from 21 to 256 and giving 236 current levels in the approximate range of 15 μ A to 1.7 mA. The assumption of excluding the first 20 current levels is substantiated by Shannon, Adams, Ferrel, Palumbo and Grandgenett (1990), where for lower current values, errors resulted in

transmitting data across the skin to the implanted electrodes. As a result, lower current levels could be disallowed to ensure that only values high enough to prevent transmission errors are utilised.

By quantising the mapped current magnitudes to the closest corresponding I_{quant} value obtained from equation 3.23, the current values that will be used for stimulation are obtained. An example of the current levels used to quantise the input current is given in figure 3.25, to illustrate the quantisation process.

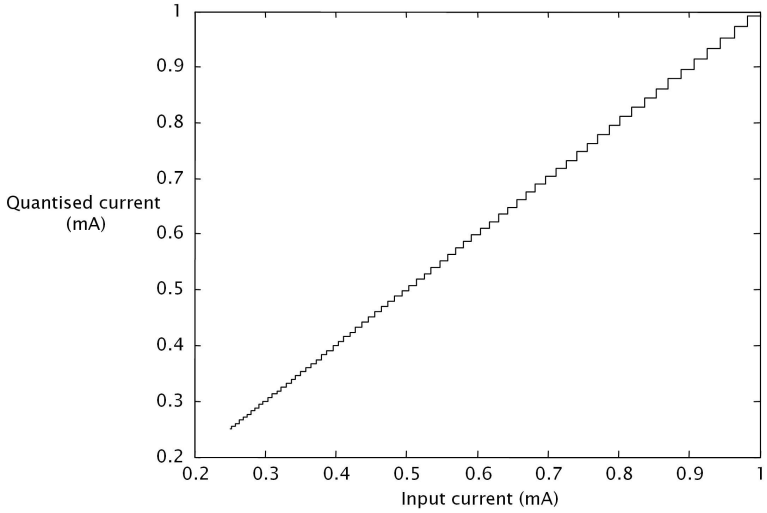

Figure 3.25.
Illustration of the 236 levels for quantisation of the input current.

An example of the resulting quantised current levels for channel 3 of the piano sound is shown in figure 3.26, using the mapped current magnitudes as shown in figure 3.24.

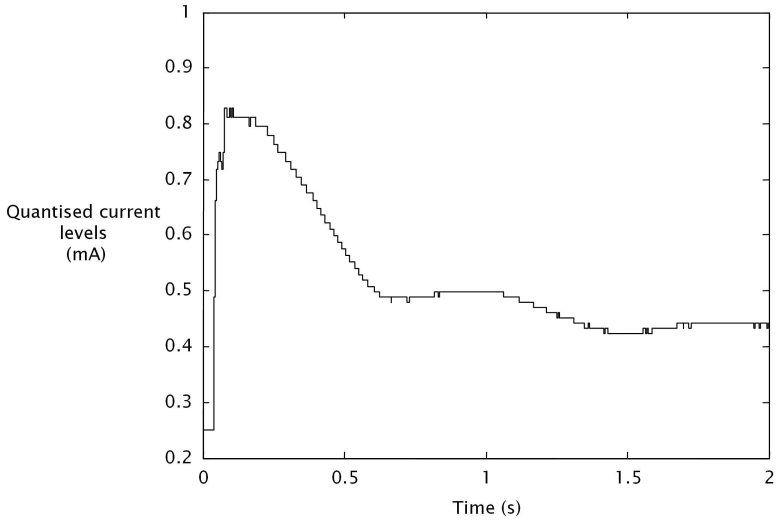

Figure 3.26.
Quantised current values for channel 3.

3.5.3 Biophysical characteristics of the acoustic model

The previous paragraphs encompass the processing part of the acoustic model. To implement a complete acoustic model, biophysical characteristics of the electrode-neural interface may be included as part of the acoustic model, as illustrated in figure 3.9. Such biophysical characteristics include the current spread along the cochlea and the shift in frequency that occurs as a result of the insertion depth of the electrode array into the cochlea. However, it was decided not to implement the biophysical characteristics of the electrode-neural interface in the final version of the acoustic model to process the musical sounds, and instead to focus solely on the effect of the processor on the sounds.

The biophysical characteristics of the electrode-neural interface provide a very generic representation of what occurs in the cochlea and in reality this differs drastically from one individual to the next. Additionally, the inclusion of the biophysical characteristics of the electrode-neural interface degraded the musical sounds to such an extent that psychoacoustic experiments would have been extremely difficult for participants, with chance responses prevailing. These factors combined with the multidimensional nature of timbre led to the decision to limit the number of parameters which could affect timbre perception through a

CI, with the aim of being able to quantitatively understand timbre in a CI processor to some extent first. Thus, to prevent the biophysical characteristics of the electrode-neural interface from obscuring effects that could be noted as a result of the processor, it was therefore decided to only include the processing part of the acoustic model.

By excluding the biophysical characteristics of the electrode-neural interface, the current spread and frequency shift are the only parts shown in figure 3.9 that are omitted, with the rest of the acoustic model remaining as is. With these sections omitted, it is implied that the quantised current values, as shown in figure 3.26, are mapped directly back to intensity values via the inverse LGF, described in section 3.5.3.1 which follows. Finally, synthesis of the acoustically modelled sound was implemented to enable a NH listener to perceive the sound acoustically, as discussed in section 3.5.3.2.

3.5.3.1 Inverse mapping to intensity values

The quantised current values must be translated back to intensity values for the purpose of constructing the synthesis signals. This is implemented by means of an inverse of the loudness growth function that was implemented in the processing step of section 3.5.2.5, to enable the current values to be mapped back to intensity values. The equation for the inverse loudness growth function can be calculated from the original loudness growth function of equation 3.22, shown in equation 3.24 below as

$$\text{intensity} = \frac{10^{I_{\text{quant}} \cdot \log(1+\alpha)} - 1}{\alpha}, \tag{3.24}$$

where I_{quant} is the calculated quantised current value for a specific electrode, and α is the same steepness factor of the curve as explained previously in section 3.5.2.5, resulting in intensity values for a given current. The intensity values can then be used as the amplitude values of the synthesis signals. Figure 3.27 illustrates the inverse loudness growth function implemented to map the quantised current values back to intensity values to be used as the amplitudes of the synthesis signals. The mapped intensity values are shown in figure 3.28.

These intensity values are used as the amplitudes of the synthesis signals, as discussed in section 3.5.3.2.

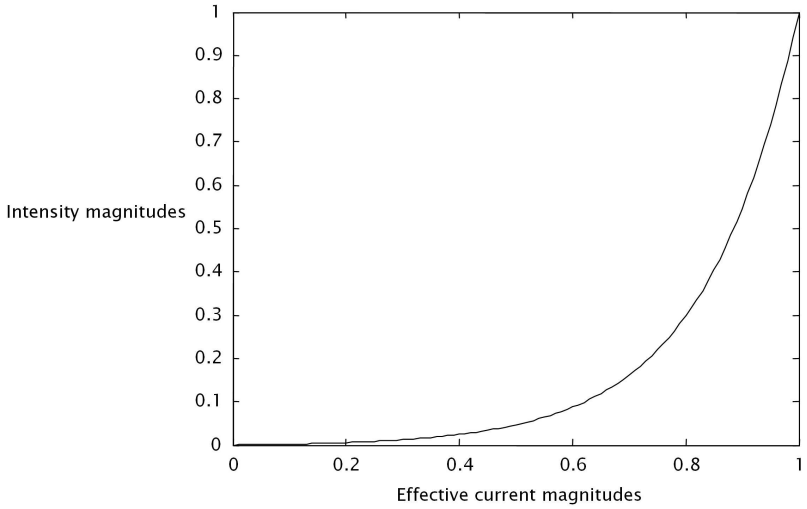

Figure 3.27.
Illustration of inverse loudness growth function to map current values to intensity values for resynthesis of the sound signal acoustically.

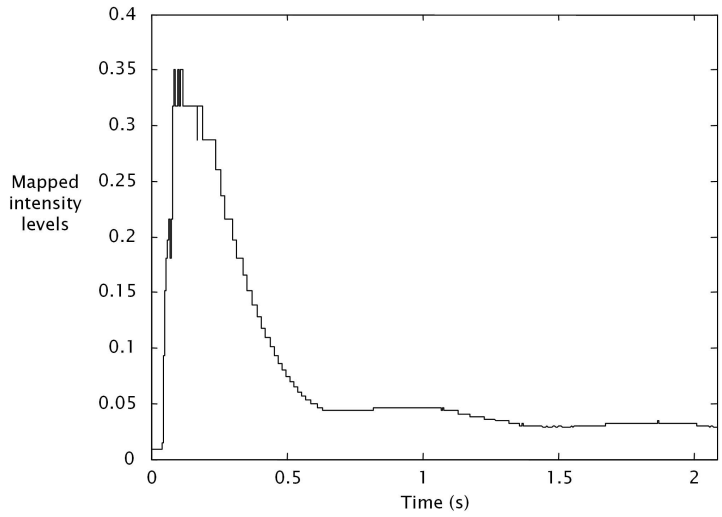

Figure 3.28.
Mapped intensity values obtained for the piano sound without the biophysical characteristics of the electrode-neural interface included.

3.5.3.2 Summation of channels to resynthesise sound

To reconstruct the instrument sound signal, sinusoids are used. Sinusoidal signals were chosen as resynthesis signals, as this modelling parameter creates a simple foundation which allows for modifications to be made, through which for example, the bandwidth of the resynthesis signal could easily be extended and the effect thereof on the reconstructed signal investigated. Sine waves are constructed for each channel for each time window, with the amplitude of the sinusoids set as the intensity values calculated as discussed in section 3.5.3.1, and the frequencies of the sinusoids corresponding to the centre frequencies of the analysis filters of table 3.1. For each time window, the sinusoids constructed for each individual channel are added together to produce an instrument sound signal with a sampling rate of the original sound signal, which can be perceived externally by a NH listener.

The resynthesised sound is normalised back to the original amplitude values between 1 and -1. Figures 3.29 and 3.30 illustrate the processed piano sound in the time and frequency domains, respectively, with the biophysical characteristics of the electrode-neural interface omitted. Even without the biophysical characteristics of the electrode-neural interface, the degradation of the signal is apparent, particularly in the frequency domain.

Examples of the acoustic modelled versions of each of the four primary musical instrument sounds, as shown in figure 3.1, are given in the time domain in figure 3.31. Figure 3.32 shows the partial representations of the four musical instrument sounds as processed through the acoustic model, which can be compared to the original musical instrument sounds shown in figure 3.3. Additionally, frequency domain representations for each of the four instrument sounds are given in figures 3.33, as processed through the acoustic model, with figure 3.2 reproduced in figure 3.34 for ease of comparison of the frequency spectra for the processed and unprocessed musical instrument sounds.

It should be noted that the resynthesised sounds as shown in figure 3.33 show the time-averaged frequency representation of the sounds. Although the selected frequency channels differed for each time window, figure 3.33 shows an average frequency representation across the duration of the resynthesised sound. The outputs from the acoustic model were scaled in amplitude between 1 and -1 to comply with .wav file specifications to be used in the experimental studies, as discussed in sections 4.2.1 and 4.2.2 that follow.

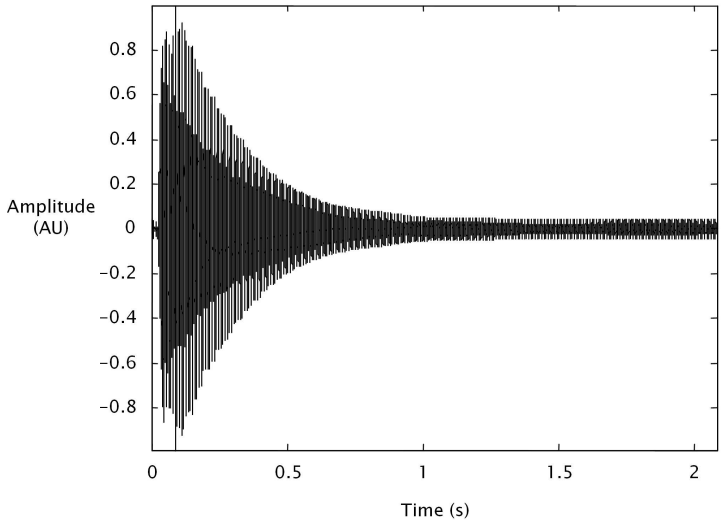

Figure 3.29.
Resynthesised version of the piano sound in time, processed through the acoustic model.

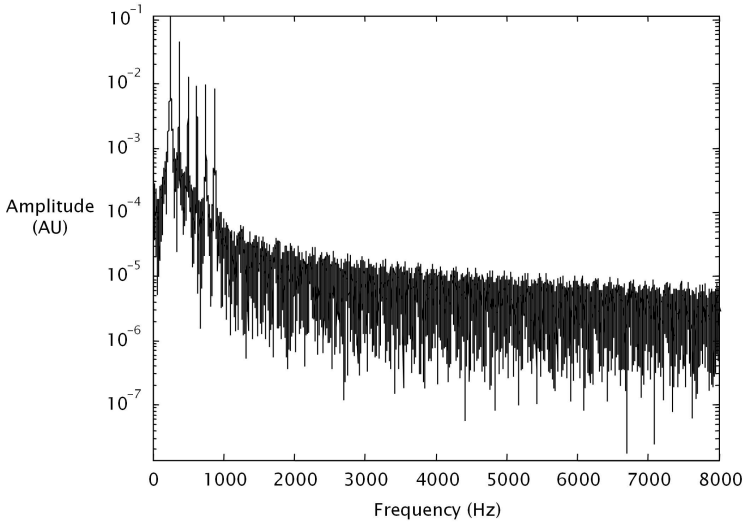

Figure 3.30.
Resynthesised version of the piano sound in the frequency domain, processed through the acoustic model.

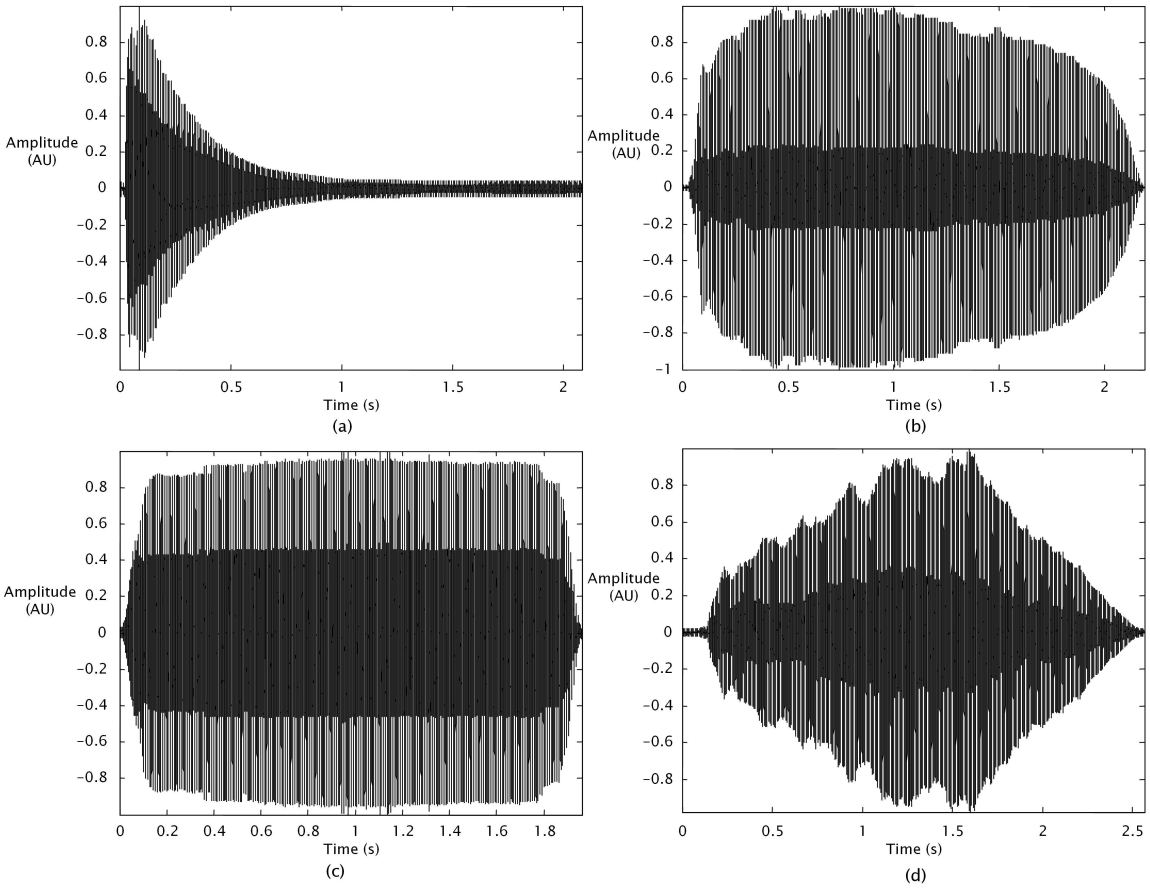

Figure 3.31.
Time domain representations for the four primary musical instrument sounds processed through the acoustic model for (a) the piano, (b) the trumpet, (c) the clarinet and (d) the violin.

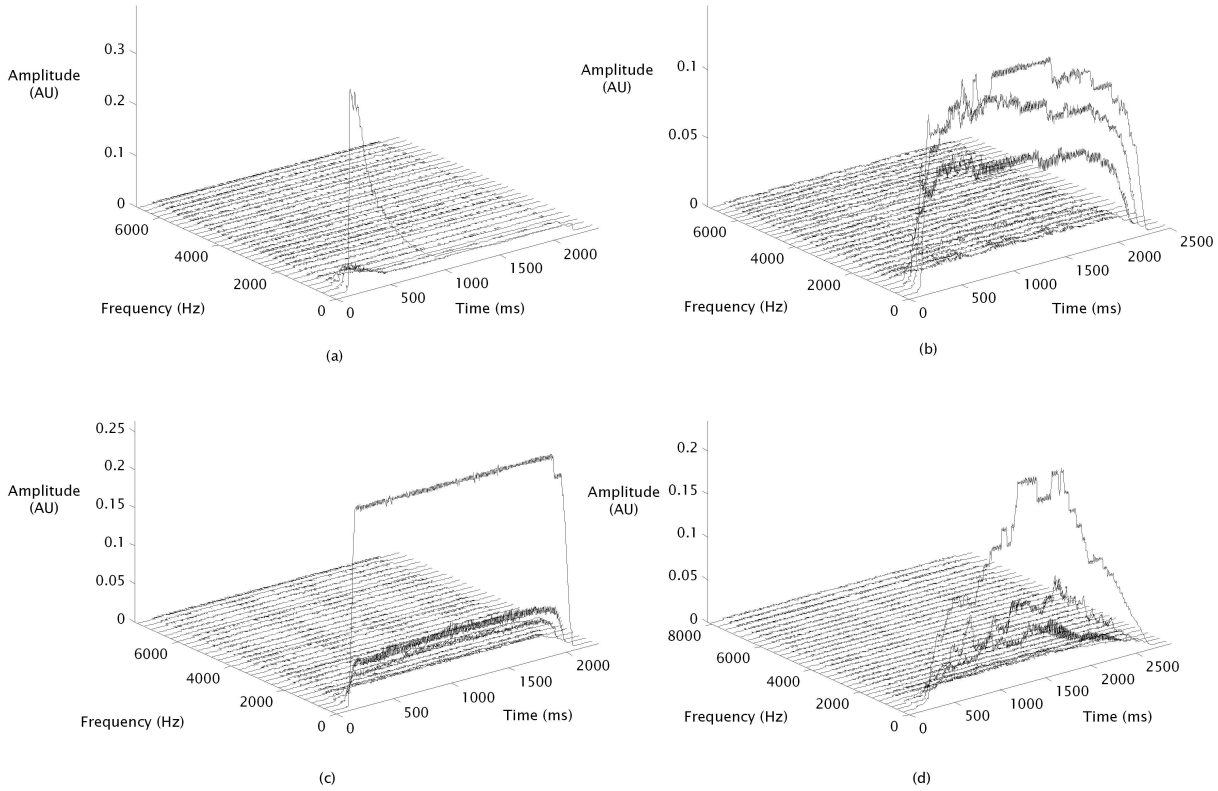

Figure 3.32.
Additive parameters for the four primary musical instrument sounds processed through the acoustic model for (a) the piano, (b) the trumpet, (c) the clarinet and (d) the violin.

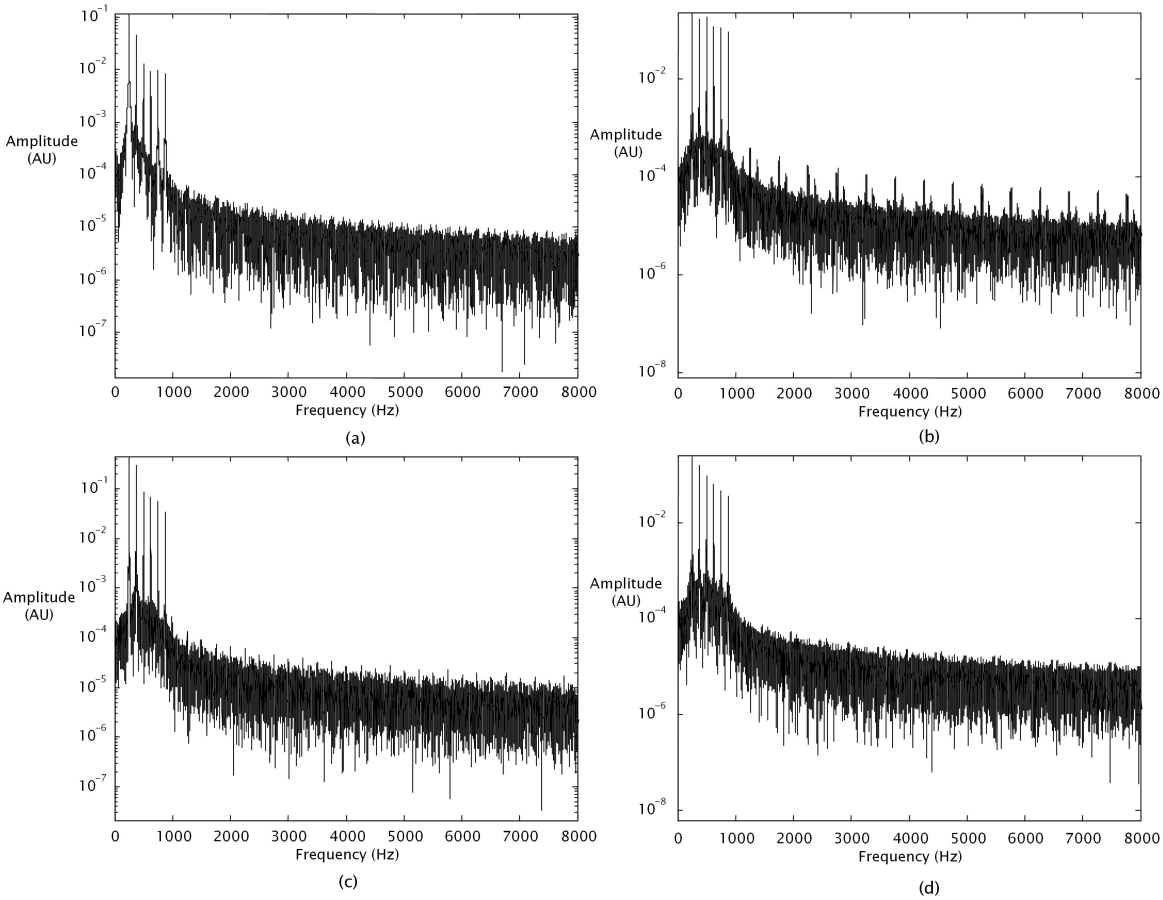

Figure 3.33.
Frequency domain representations for the four primary musical instrument sounds processed through the acoustic model for (a) the piano, (b) the trumpet, (c) the clarinet and (d) the violin.

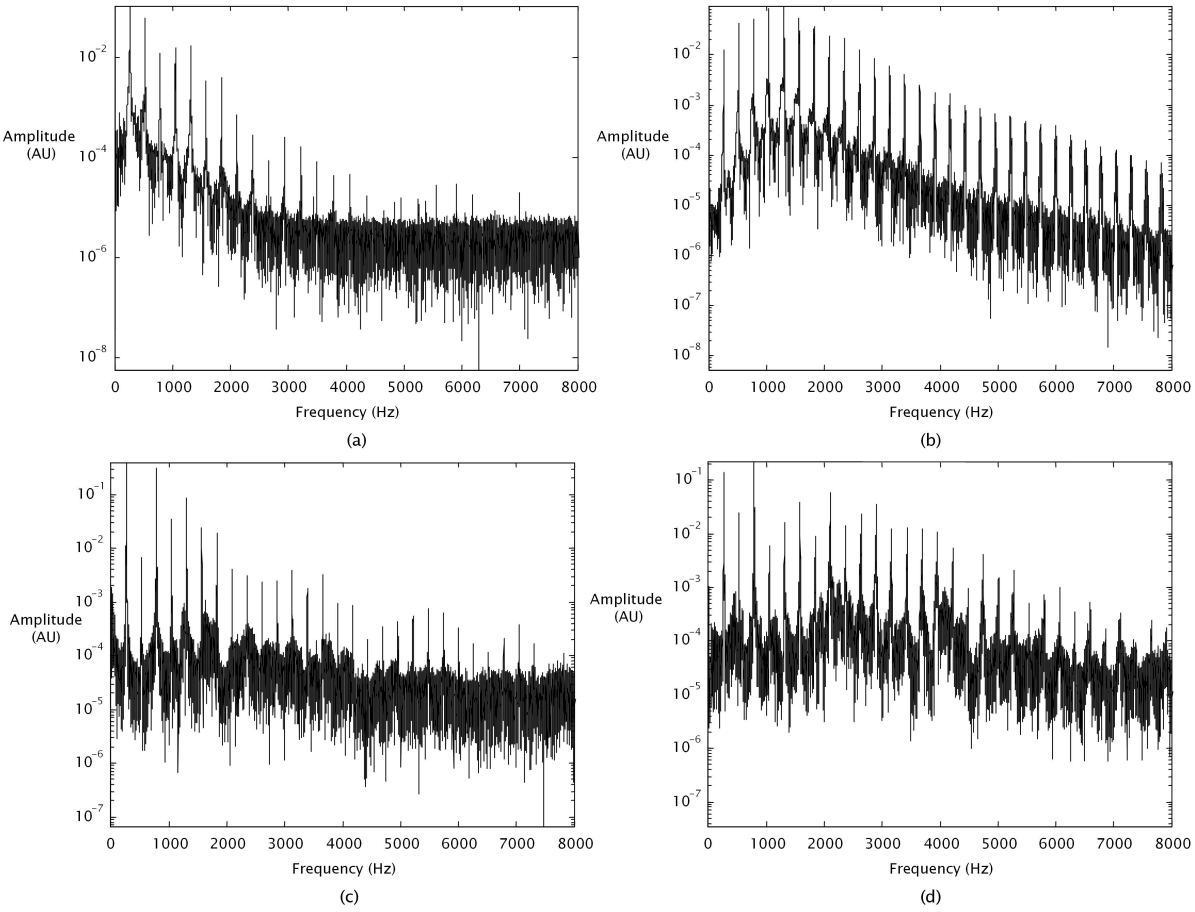

Figure 3.34.
Frequency domain representations for the four primary musical instrument sounds as given in figure 3.2 for (a) the piano, (b) the trumpet, (c) the clarinet and (d) the violin.

3.5.3.3 Discussion of acoustic model effects

By comparing figures 3.33 and 3.34, it can be seen that the frequency domain representations of the musical instrument sounds differ substantially for the processed and unprocessed sounds. The frequency domain representations of each of the unprocessed musical instrument sounds in figure 3.34 are clearly distinct from one another, as can be seen by the shapes of the frequency spectra. For example, the piano sound 3.34(a) has peaks only in the lower frequency ranges, while the trumpet sound 3.34(b) has high frequency values across the spectrum into much higher frequency values. The frequency peaks of the clarinet sound 3.34(c) and violin sound 3.34(d) take on a more irregular spectral shape. However, in the case of the processed musical instrument sounds in figure 3.33, the appearance of the frequency spectra of the four sounds are very similar. This would be expected due to the processing step of the acoustic model where energy calculations of the frequency bands over time windows are performed. This processing step is indicated in the second step of the processor model in figure 3.9, and is discussed in section 3.5.2.2. As a result of this processing step, both the fine temporal structure and fine time structure of the sounds are destroyed, leaving the processed sounds with similar spectra.

Figure 3.33 also shows that, for an average across the duration of each sound, the lowest frequency components of the sound are selected as the stimulating channels, as can be seen from the low pass appearance of each of the sounds, with distinct peaks clustered in the lower frequency ranges. This is as a result of the peak-picking algorithm employed by the ACE processing strategy, where the strongest frequency components are selected as the stimulating channels. In the case of this study, this is always synonymous with the components in the lower frequency ranges being selected. It should be noted that this result could be different if pre-emphasis of the sounds was performed prior to the processing steps implemented in the acoustic model for this study. Pre-emphasis may cause stronger frequency components in the higher frequency ranges to be selected as the stimulating channels.

The drastic effect of the implemented acoustic model on the musical instrument sounds is evident, and this effect is only as a result of the CI processor. Even without the biophysical characteristics of the electrode-neural interface, such as current spread, the acoustic model implementation of the processor greatly affects the frequency spectra of the musical instrument sounds. Illustrations of the other 6 instruments as processed through the acoustic model are given in Appendix A.

3.6 SUMMARY

This chapter provided detail regarding the approach followed to investigate timbre perception in the electrically stimulated auditory system. The database of musical instrument sounds used in this study was presented, as well as the extraction of important timbre features from these sounds. A full description of the acoustic model implementation was given, based on knowledge of CI characteristics and existing acoustic models. Design considerations of the acoustic model were also discussed. The methods presented in this chapter provide a foundation on which to develop both experimental procedures and a model of timbre perception, which will be discussed in detail in chapters 4 and 5, respectively.

CHAPTER 4

MEASUREMENT OF TIMBRE PERCEPTION

A large part of the work described in this chapter was presented at the CI 2010 conference (Hanekom and Hugo, 2010) and has also been submitted to Ear and Hearing in the form of a journal article.

4.1 CHAPTER OBJECTIVES

Using the methods presented in Chapter 3 as a basis, the experimental component of this study could be developed, consisting of two different experimental procedures. The first experiment was developed to measure timbre perception in NH and CI listeners by means of discrimination tasks, the results of which were used in the model of timbre perception. The second experiment consisted of similarity ratings of musical instrument timbres and were used to validate the outcomes predicted by the model of timbre perception. This chapter presents the experimental procedures, which were developed in Matlab version 2007b, as well as the results obtained from these experiments. For the first experiment, JNDs were found for B, LRT and IRR and are presented in this chapter for NH and CI listeners. Following this, results of the second experimental study of the similarity ratings of musical instrument timbres are presented. A discussion of the experimental results concludes the chapter, with comparisons of the results with existing literature explored in detail.

4.2 METHODS

4.2.1 Discriminations of timbre perception features

The first experimental study was performed to measure discrimination abilities of listeners for each of the three important timbre features: B, LRT and IRR. Discrimination tasks were carried out for pairs of synthesised sounds, where the only differences between the sounds each time were variations in the value of the timbre feature being investigated. This experimental study was carried out for NH and CI listeners as the first step in quantitatively understanding timbre perception differences for these two groups. Additionally, the results of this study were used in the implementation of the model of timbre perception (see section 5.2.1) to predict the outcomes of the experimental results documented in section 4.2.2. Details of the experimental procedure are discussed in the following paragraphs and the results are presented in section 4.3.1.

4.2.1.1 Listeners

Five NH and five CI listeners participated in the study. Participants were age-balanced across the two groups. The five NH listeners (two females, three males), were aged between 24 and 66 years (average age = 39 years). Each listener was screened to ensure that the criterion for normal hearing was adhered to. Normal hearing was defined as achieving audiometric thresholds of 30 dB HL or better over six octaves (250 to 8000 Hz), and all five listeners were confirmed as NH participants. None of the NH participants had any formal music training.

The five postlingually deafened adults (four females, one male), were aged between 21 and 66 years (average age = 42.4 years). CI listeners all used the Freedom processor, and had three or more years' experience with the implant system. Four CI participants used the ACE processing strategy, and one (CI 2) used the SPEAK processing strategy. Only one of the CI listeners had been exposed to formal musical training (CI 5), but was not actively studying music or playing an instrument at the time of the experiments. Three of the subjects had previously participated in CI studies performed in our research group. Additional relevant information regarding the CI listeners is given table 4.1. For the type of strategy used, the stimulation rate (SR) in Hz for each channel is indicated in brackets. Listeners with an asterisk marked next to the implanted ear have bilateral implants.

Table 4.1.
Details of CI subject demographics.

Subject	Sex	Age	Processor	Implant	Strategy (SR per channel)	Years implanted	Test ear
CI1	F	59	Freedom	24R (CA)	ACE (500 Hz)	4	Right*
CI2	F	21	Freedom	Nucleus 22 Series	SPEAK (250 Hz)	>10	Left*
CI3	M	66	Freedom	Freedom (CA)	ACE (1200 Hz)	4	Right*
CI4	F	44	Freedom	Freedom (CA)	ACE (900 Hz)	3	Right
CI5	F	24	Freedom	24R (CA)	ACE (900 Hz)	5	Left

All 10 listeners gave written informed consent for their participation before commencing with the study, based on guidelines presented by the appropriate ethics committee. The listeners were compensated for their time upon completion of the experimental sessions.

4.2.1.2 Stimuli

A tone of two seconds, consisting of 30 harmonics, was created by additive synthesis in Matlab version 2007b, with a sampling frequency of 16 kHz. The fundamental frequency was chosen as 262 Hz (C4 or middle C), a common note used in Western music and therefore in timbre studies such as those by Gfeller *et al.* (1998) and Nimmons *et al.* (2008). The stimulus was varied along the three timbre features B, LRT and IRR, and the features never co-varied. The synthesised sound was then adjusted by altering the value of one of the three timbre features.

The tone always consisted of five linear segments: a start segment, followed by a rise time segment, a plateau, a release and an end segment, to complete the 2 s length of the sound (see section 2.3.3 for a full description of the amplitude envelope model). The spectrum was harmonic, where the amplitude spectrum, A, at each point in time was a function of B, and

the rank of the harmonic, k, and is given by equation 4.1 as

$$A(k) = \left(\frac{\mathrm{B}}{\mathrm{B}-1}\right)^{-k}. \tag{4.1}$$

The length of the rise time segment of the synthesised tone was constructed using LRT each time, as given by equation 3.17 in section 3.4. The starting point in time of the rise time segment was fixed, indicating that a change in LRT would simply change the point in time that the sound reached a maximum amplitude. To implement irregularity in the spectrum in the synthesised tone by using a given IRR value, a deviation, d, from the original amplitude values found from the required B value was calculated as in equation 4.2 by

$$d = \frac{10^{IRR}}{N-2}, \tag{4.2}$$

with N being the total number of harmonics (30 in this study). The deviation value, d, was then added to the log of the amplitudes of all of the even harmonics and subtracted from the log of the amplitudes of the odd harmonics of the synthesised sound.

The reference tone had a start segment of 140 ms, a rise time of approximately 316.22 ms (corresponding to the selected reference LRT of 2.5), a sustain or plateau of 600 ms, a release time of 300 ms, and an end segment for the remainder of the total 2 seconds of 643.77 ms. The B value of the reference synthesised tone was set at 4 and the IRR value was set at zero. Figure 4.1 illustrates the reference tone used in this study.

In the experiment, each of the three features (B, LRT and IRR) was tested. In each case, a reference tone was presented in conjunction with a tone which had been altered from the reference tone by adjusting the feature under investigation. The reference tone was slightly different for each feature being investigated, so that the altered tone could be varied for parameter values that were both lower and higher than the reference tone.

When B was varied, the reference tone was set to have a B value of 4, an LRT of 2.5 and an IRR of 0. The initial B values of the tones which were to be compared to the reference tones were set at 6.2 and 1.8. Figure 4.2 illustrates the variations in spectral components for these extreme initial values of B.

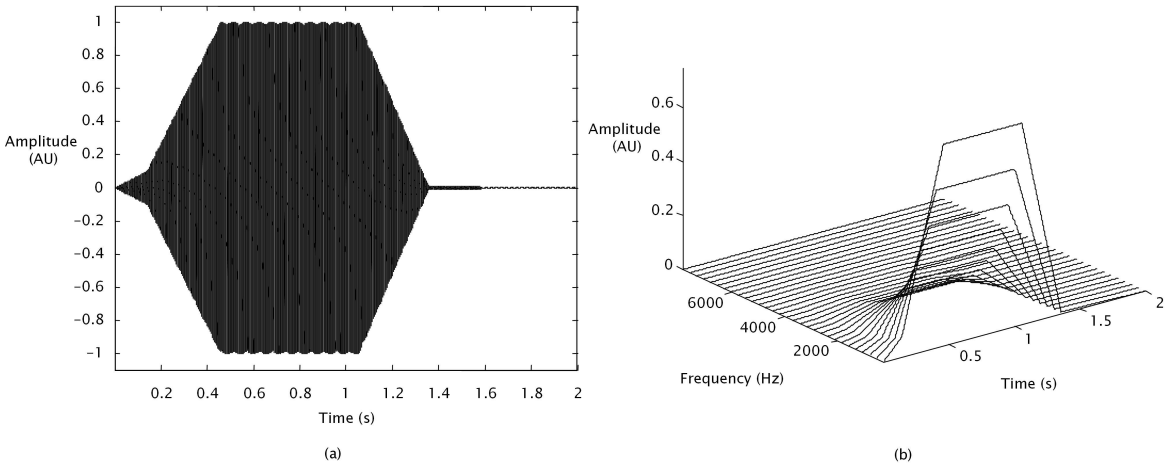

Figure 4.1.
Reference synthesised tone used for the first experimental study illustrated in (a) the time domain and (b) in terms of additive parameters, with B = 4, LRT = 2.5 and IRR = 0.

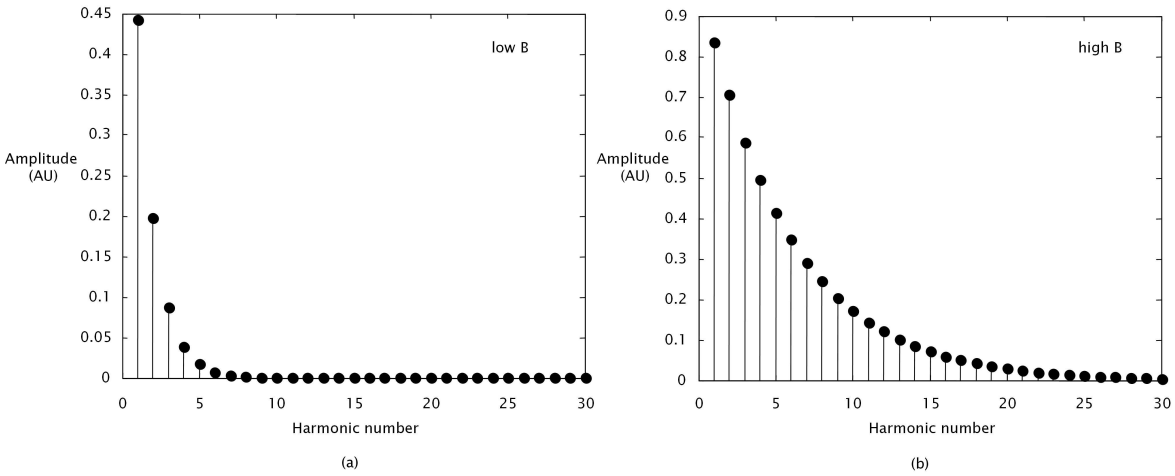

Figure 4.2.
Illustration of variations in B for (a) a low B value of 1.8 and (b) a high B value of 6.2.

When LRT was varied, the reference tone had a B value of 4 and an IRR value of 0, with the default LRT set at 1.8. The initial LRT values of the tones which were to be compared with the reference tones were 0.5 and 2.9. Figure 4.3 illustrates the variations in the amplitude envelope for these extreme initial values of LRT.

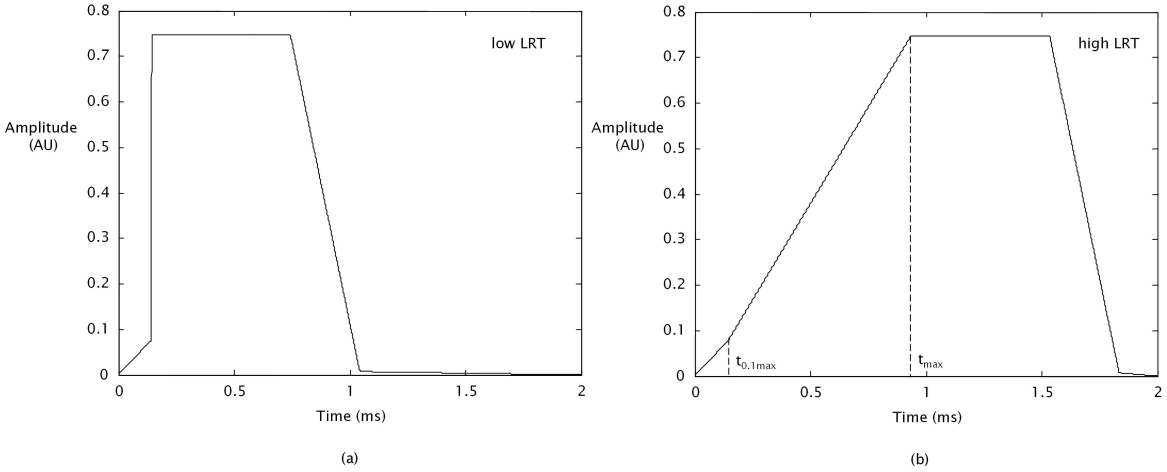

Figure 4.3.
Illustration of variations in LRT for (a) a low LRT value of 0.5 and (b) a high LRT value of 2.9. The time values from which LRT is calculated (see equation 3.17) are shown in (b).

When IRR was varied, the reference tone had a B value of 4, an LRT value of 2.5 and an IRR value of 2. The initial IRR values of the tones which were to be compared with the reference tones were 0.1 and 4. Figure 4.4 illustrates the variations in the spectral components of the sound for different IRR values.

The stimuli were presented in sound field and at the same perceived loudness level for all subjects. A loudness estimation procedure was used, whereby 1 kHz tones ranging between the lowest and highest comfortable loudness levels for each subject were scaled to find individual sets of intensity levels. The tones were presented 20 times each at 10 linearly spaced intensity levels, to find an average estimated perception of loudness at each level. These sets were then interpolated to find the intensity level corresponding to 50 % of the subject's perceived loudness level, at which the stimuli in the experiment were presented (at 50 % of the maximum comfort level).

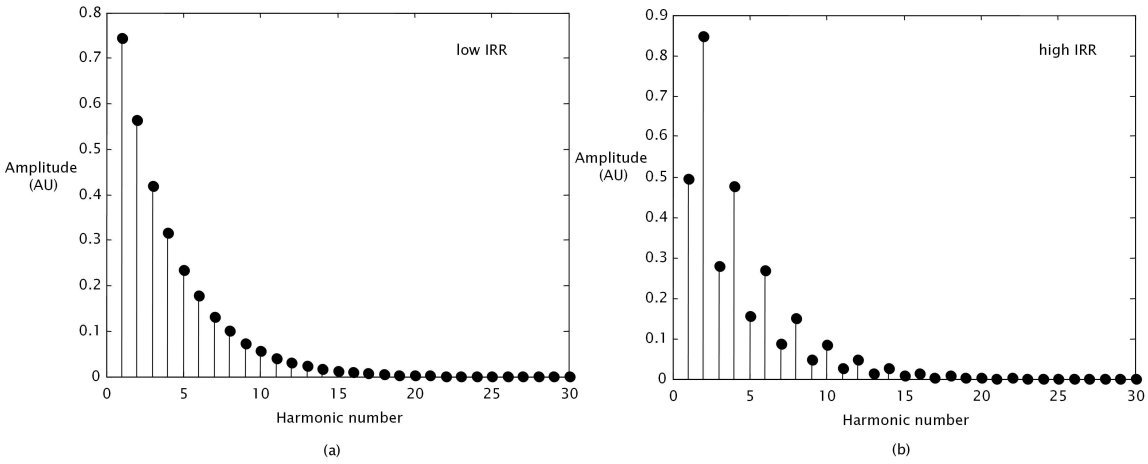

Figure 4.4.
Illustration of variations in IRR for (a) a low IRR value of 0.1 and (b) a higher IRR value of 2.

4.2.1.3 Procedure

Experimental sessions were conducted in a double-walled sound booth. The experimental procedure was controlled using software, with programming done in Matlab version 2007b. Sounds were presented from the computer via an external M-Audio Fasttrack Pro audio interface (44.1 kHz, 16 bits), with a Yamaha MS101 II speaker positioned approximately 1 m in front of the subject.

Timbre feature JNDs were determined using an adaptive two-alternative forced choice (2AFC) procedure. Each trial consisted of two synthesised tones, each two seconds long, separated by an interstimulus gap of 200 ms, with one tone always corresponding to the reference tone for the timbre feature under investigation, and the other tone the altered sound with adjusted timbre feature. The subject was asked to decide whether the two tones sounded the same or different, by choosing either of two buttons labelled “same” and “different” on the screen. Subjects were only allowed to listen to the tone pair once and were not provided with feedback. The next tone pair was only presented once the listener response had been made, allowing the subject adequate time to make a decision. A screen shot of the experimental graphical user interface (GUI) is shown in figure 4.5.

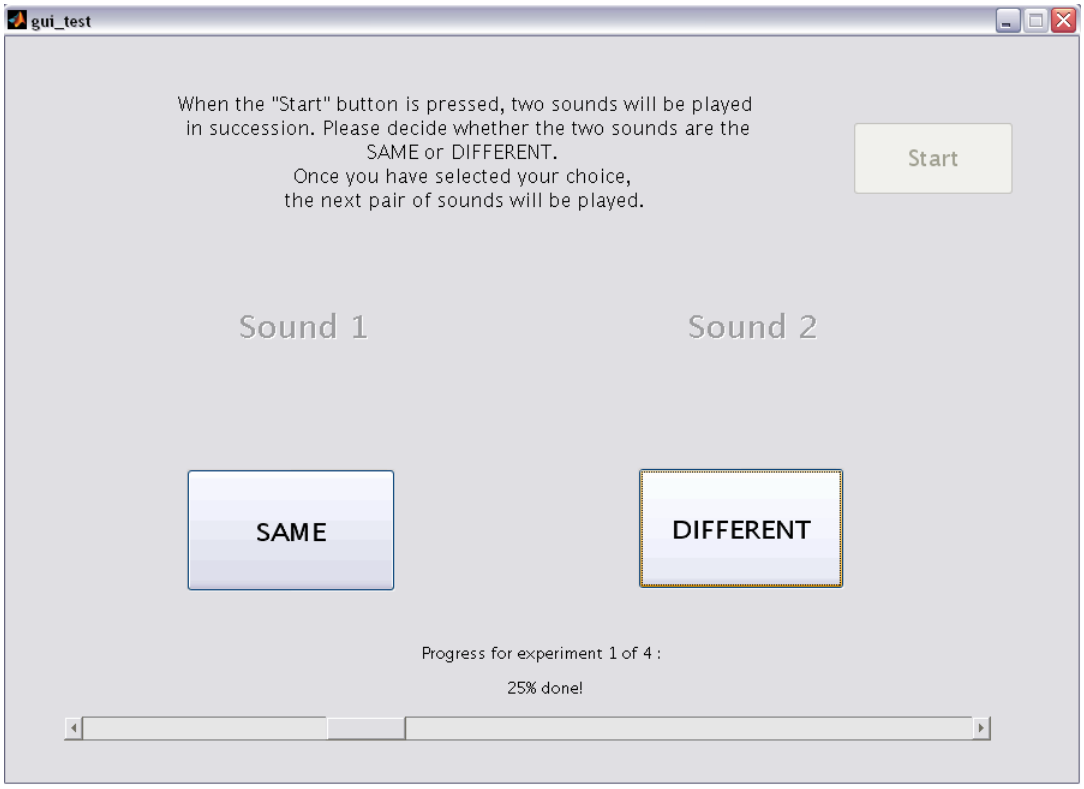

Figure 4.5.
Illustration of the GUI used for the timbre feature discrimination experiments.

Each experimental session consisted of four repeats of the experiment for each timbre feature investigated. The subjects were unaware that the first experiment was considered a practice session, to familiarise the listener with the task and sounds presented. A progress indicator was included in the user interface to assist the listener in gauging their progress for the experimental session.

An adaptive procedure based on the transformed up-down staircase technique (Levitt, 1971) was implemented, using a 2-down, 1-up decision criteria. The starting values of the altered tones, or probe tones, compared to the reference tone, were described in section 4.2.1.2 and adjusted accordingly with the staircase method as a result of the listener's response. Two interleaved adaptive procedures were implemented for each experiment: one with the probe tone starting at a higher parameter value than that of the reference, and one with the probe tone starting at a lower parameter value than that of the reference (Jesteadt, 1980). These sequences were presented in a random order, as were the reference and probe tones for each trial. The technique employed allows the probe stimulus to alternate between values that make the sound just distinguishable and just not distinguishable from the reference tone. If the difference could not be detected, the difference between the feature value of the reference and probe tone was increased until the listener could again detect the difference. One such oscillation in the response is classified as a reversal. Two consecutive correct responses to the stimulus pair resulted in the difference between the timbre feature for the probe and reference tones being reduced, while one incorrect response resulted in an increase in the difference. The value to which this difference converged was accepted as the JND in each case.

A total of 10 reversals were recorded for each of the two interleaved sequences. For the first two reversals, an adaptive factor of 1.6 was used, while the remaining eight reversals were subjected to an adaptive factor of 1.2. The JND was calculated from the average of the midpoints of the last five reversals for each of the two interleaved sequences, to find an average JND for the experimental session. A final JND value for each timbre feature was then calculated from the average over three repeats of the experiment for each listener.

Although subjects were informed about the nature and aim of the study before commencing with the experiments, they were not aware of the procedure used in presenting the stimuli. Due to the varying availability of subjects, experimental sessions were completed over the course of several weeks.

4.2.2 Similarity ratings of timbre

The second experimental study consisted of similarity ratings of different musical timbres. The results of this study, carried out for both NH and CI listeners, served as experimental timbre perception data to which the predictions from the model of timbre perception, as discussed in section 5.2.1, could be compared. The similarity rating experiment is explained in detail in sections 4.2.2.1 to 4.2.2.3. This study served as a basis for validation of the model of timbre perception that was developed, as well as for the implementation of the acoustic model, to which the NH participants were exposed.

A similarity rating experiment was set up to determine the confusions between pairs of musical instrument sounds. Similarity ratings were chosen because direct identification tasks would be impractical for CI listeners, as many would not have the musical memory to accurately identify a range of different instruments. Since the same experiments were required to be carried out on both NH and implant listeners, the similarity judgement was the most feasible approach and was based on the technique used by Getty, Swets, Swets and Green (1979). This method, further explored in Getty, Swets and Swets (1980), makes use of similarity judgement experiments, which are then used to perform MDS of the perceptual dimensions. The distances obtained from the scaling are then related to confusions between stimuli by means of an identification task. Using this approach as a basis, the steps followed in this study are used to perform similarity judgement tasks directly. The similarities can then be expressed directly as percentages of confusions between sounds, as in the case of a standard confusion matrix constructed from identification tasks (section 5.2.2 provides more information on confusion matrices). These results can then be directly compared to the predictions made from the model of timbre perception.

4.2.2.1 Listeners

The participants for the similarity rating experiment were the same listeners that participated in the first experiment as in section 4.2.1, and consisted of five NH and five post-lingually deafened adults.

4.2.2.2 Stimuli

The experimental session involved the comparison of the timbres of pairs of the 10 musical instrument sounds as described in section 3.2, by rating their similarity. The instrument sounds were taken from the sound database of Fritts (No date), and each of the instrument sounds was approximately 2 s long, ranging from a minimum of 1.98 seconds to a maximum of 2.6 seconds. All of the samples were recorded in mono with a sampling rate of 44.1 kHz (16 bit) and were saved in .aiff format. The only exception was the piano, which was recorded in stereo. The University of Iowa Musical Instrument Samples website (Fritts, No date) contains all the relevant information regarding the instrument sound recordings.

For this study, sounds with a fundamental frequency of 262 Hz (C4 or Middle C) were used, as discussed in section 3.2. The sounds were normalised in amplitude between 1 and -1 to comply with .wav file specifications and to ensure that the sounds were presented at the same amplitudes. The saved .wav files of the instrument sounds, while not explicitly controlled in duration, were chosen to represent orchestral families of instruments (Galvin III *et al.*, 2008), and listeners were instructed to focus solely on timbre differences. For CI listeners, the original recorded sounds were presented in pairs as the stimuli, while for NH listeners, the stimuli were the instrument sounds processed through the acoustic model, as discussed in section 3.5. A pilot experimental study for the NH listeners subjected to the orignal recorded musical instrument sounds was also carried out. The stimuli were presented in sound field and at a comfortable perceived loudness level for each of the subjects, between 50 % and 70 % of the maximum comfortable loudness level in all cases.

4.2.2.3 Procedure

The routine for the experiment was programmed in Matlab (2007b), with experimental sessions conducted in the same way as for the first experimental procedure described in section 4.2.1.3, in a sound proof booth with a Yamaha MS101 II speaker positioned approximately 1 m in front of the subject. Subjects received on-screen written and verbal instructions for the experimental task and clarification was given if required. The 10 instrument sounds, unprocessed for CI listeners and processed for NH listeners, were presented in a random order to familiarise the listener with the range of variation amongst the timbres that were to be rated on a 10 point scale. The subject could listen to all of the sounds a maximum of three times,

or less if they were comfortable with the set of stimuli. For each experimental trial, a pair of instrument sounds was presented, with a pause of 2.9 seconds between the start of the first sound and the start of the second. The listener was asked to rate the degree of dissimilarity between the sounds on a scale of 0 (exactly the same) to 9 (very different). The sound pair could be repeated up to four times before a rating had to be submitted via the user interface by selecting one of the buttons labelled 0 to 9. Figure 4.6 illustrates the GUI for the similarity rating experiment.

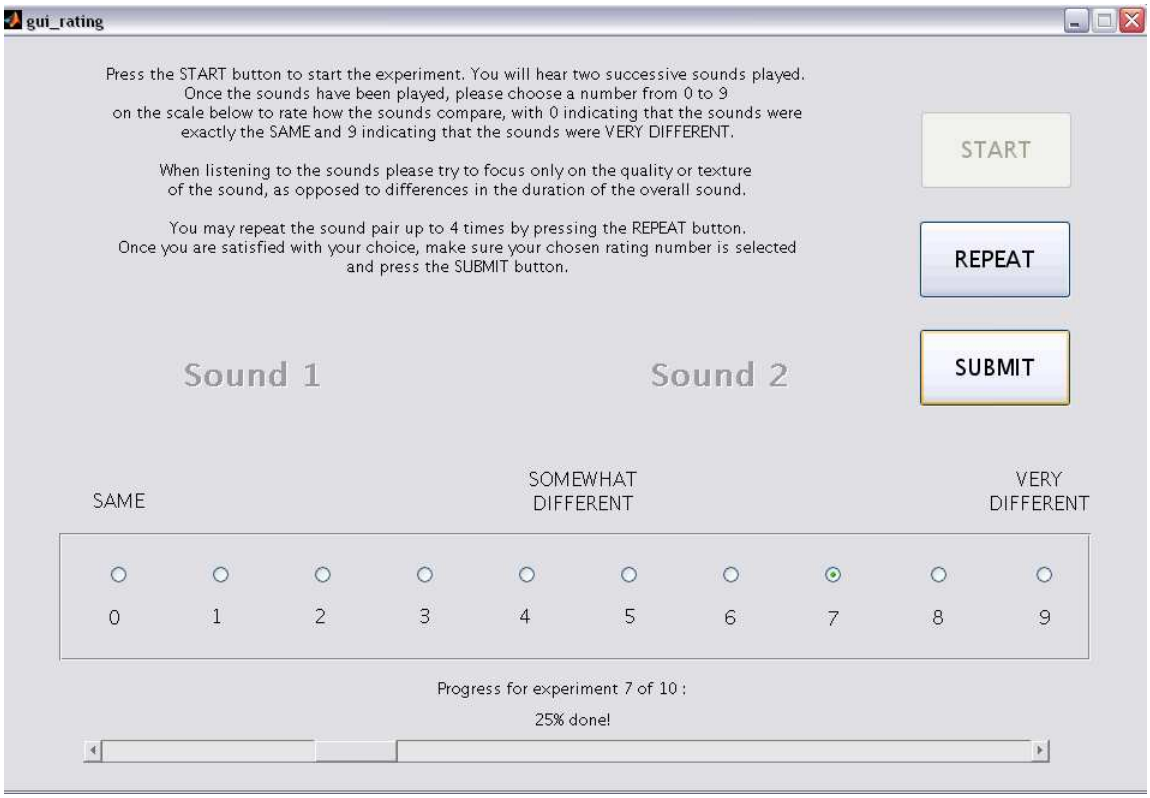

Figure 4.6.
Illustration of the GUI for the musical instrument timbre similarity rating experiment.

Subjects were requested to use the entire scale when making their decisions. In total, 100 sound pairs (all the possible combinations of the 10 sounds) were presented in a random order in one experimental run. A total of 10 experimental runs, each approximately 20-30 minutes in length, was conducted for each listener, to obtain an average dissimilarity rating for each listener for each sound pair. Learning effects were observed in the results of the rating experiments for each subject. However, no noticeable trends were found and

therefore an average of all 10 experimental runs for each subject was used in finding an average confusion matrix. An additional similarity rating experiment was conducted for the NH listeners in response to the original or unprocessed musical instrument sounds. However, due to time constraints, only three experimental runs were conducted for each listener, to find an average confusion matrix. These results were used to illustrate NH subject performance in NH conditions.

The experiments were again completed over several weeks, according to the availability of the subjects. The similarity judgements obtained from the experimental sessions for each subject were converted to confusion matrices to be analysed and compared to the confusion matrices as predicted by the model of timbre perception. The development of the timbre perception model to predict the outcomes of this experimental study is discussed in section 5.2.1.

4.3 RESULTS

4.3.1 Results of timbre feature discriminations

Figure 4.7 shows the results of the first experimental study, described in section 4.2.1. JNDs obtained from the discrimination tasks for each of the timbre features B, LRT and IRR are shown for individual subjects for both NH and CI listeners, with the mean and SD values given in table 4.2.

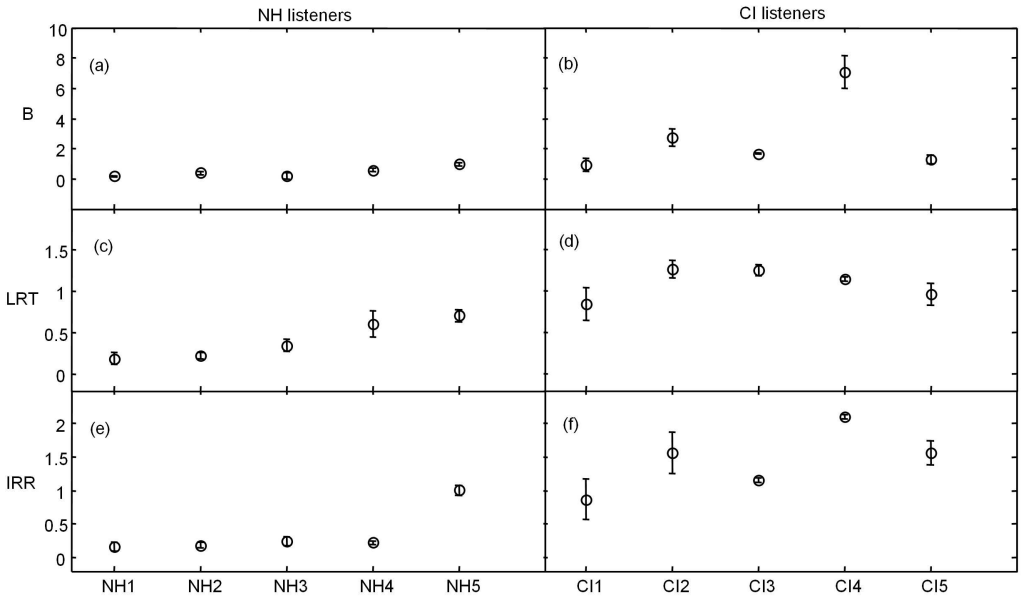

Figure 4.7.
Results of the timbre feature discrimination tasks, with the mean and SDs of the JNDs for each listener given for features B for (a) NH and (b) CI listeners, LRT for (c) NH and (d) CI listeners, and IRR for (e) NH and (f) CI listeners. The mean values are indicated by the unfilled circles, while the SD values are illustrated by the errorbars for each subject, with the units for B, LRT and IRR given in table 4.2.

Table 4.2.
Mean and SD values for the JNDs obtained for the timbre feature discrimination tasks, with the units for B, LRT and IRR given by partial index, log(s) and log(dB), respectively.

		NH listeners					**CI listeners**				
		NH1	NH2	NH3	NH4	NH5	CI1	CI2	CI3	CI4	CI5
B	Mean	0.177	0.393	0.213	0.602	0.987	0.963	2.714	1.684	7.058	1.309
	SD	0.047	0.085	0.227	0.098	0.103	0.435	0.572	0.067	1.091	0.273
LRT	Mean	0.194	0.226	0.353	0.609	0.714	0.851	1.274	1.256	1.147	0.964
	SD	0.077	0.037	0.070	0.154	0.071	0.196	0.105	0.067	0.027	0.130
IRR	Mean	0.163	0.190	0.247	0.223	1.011	0.869	1.566	1.154	2.104	1.570
	SD	0.070	0.044	0.069	0.028	0.074	0.301	0.311	0.029	0.028	0.178

Figure 4.8 shows the pooled JND values obtained for each of the features B, LRT and IRR. The NH listener JNDs were averaged for each feature, as were those obtained by the CI listeners, with the mean and SD values shown in table 4.3.

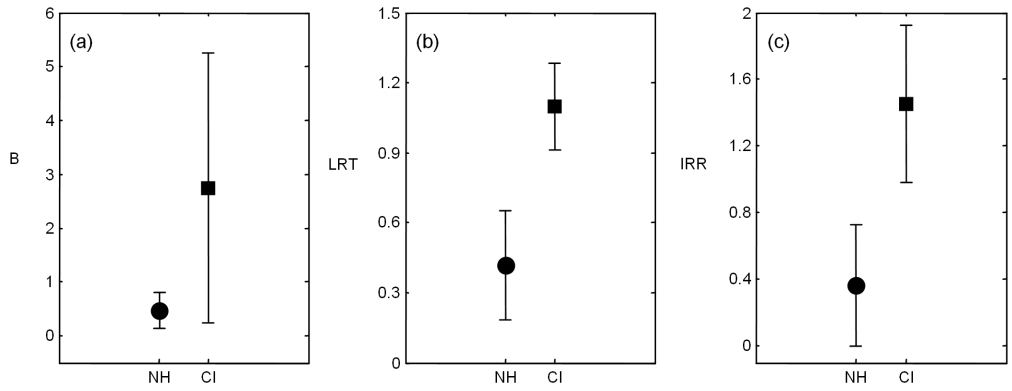

Figure 4.8.
Average results for timbre feature perception for (a) B, (b) LRT and (c) IRR for NH listeners (circles) and CI listeners (squares). The SD values of the group for each feature are displayed by the errorbar, with the units for B, LRT and IRR as defined in table 4.3.

Table 4.3.
Averaged mean and SD values for the JNDs obtained for the timbre feature discrimination tasks, with the units for B, LRT and IRR given by partial index, log(s) and log(dB), respectively.

	NH listeners		**CI listeners**	
	Mean	SD	Mean	SD
B	0.4742	0.3325	2.7453	2.4985
LRT	0.4193	0.2319	1.0986	0.1853
IRR	0.3668	0.3613	1.4526	0.4690

Figure 4.8 shows that the mean B JNDs for all NH listeners was lower than that of the CI listeners. Results of NH listeners (mean = 0.4742, SD = 0.3325) were also more consistent than those of the CI listeners (mean = 2.7453, SD = 2.4985), with variations in mean and SD values amongst listeners being substantial. The LRT JNDs for all NH listeners compared to CI listeners were also lower (figure 4.7(c) and (d)), but with less variation amongst listeners than in the case of the B JNDs. Consequently, although the mean JND of the NH listeners of 0.4193 was again lower than that of the CI listeners of 1.0986, the SDs (NH = 0.2319, CI = 0.18525) were comparable, with the SD of all the CI listeners even lower than that of all NH subjects. For the IRR JND values, figure 4.7(e) and (f) show that the mean JNDs for NH listeners were again lower than for CI listeners, with the exception of one NH listener. The mean of 0.3668 for all the NH listeners was again lower than the mean of 1.4526 for all the CI listeners. The JND SDs of 0.469 and 0.3613 for NH and CI subjects, respectively, were comparable for IRR.

The effect of the listener type (NH and CI) and the effect of the specific timbre feature (B, LRT and IRR) on the resulting JNDs were investigated by means of a two-factor analysis of variance (ANOVA). Levene's test for equal variances showed a significant result ($F(5,24) = 4.16$, $p < 0.05$), indicating that the assumption of non-equal variances should be used. The analysis revealed that the JND values for B, LRT and IRR were significantly different for NH and CI listeners ($F(1,24) = 11.993$, $p < 0.05$). This was expected due to the poor timbre perception abilities of CI listeners, as reported in the literature. The JNDs were not significantly affected by the timbre feature, with JNDs not significantly different for any comparisons of the timbre features, B, LRT and IRR (in all cases $p > 0.05$). There was also a non-significant interaction effect between the type of listener, NH or CI, and the timbre feature on the JND ($F(2,24) = 1.511$, $p > 0.05$). This indicates that NH and CI listeners were not affected differently by different timbre features. The highest mean JND values for both NH and CI listeners were for B. The lowest mean JND value for NH listeners was for IRR, while for CI listeners LRT was the feature with the lowest mean JND.

4.3.2 Results of timbre similarity ratings

The timbre similarity rating experiments explained in section 4.2.2 can be expressed directly as percentages of confusions between sounds, as the more similar two sounds are, the higher the probability of confusing the two will be. For example, if two sounds were rated as 0 (no difference) on the ten-point scale used in the similarity ratings, this would indicate a similarity of 100 %, or 100 % confusion, whereas a rating of 9 (very different), would correspond to 0 % similarity and thus 0 % confusion between the two sounds. For a pair of sounds rated at 7, this would correspond to a probability of confusion of 0.22 between the sounds (a small chance of confusing two sounds that are perceptually quite different). Each row of the matrix is normalised to the sum of that row. The average confusion matrices obtained for NH and CI listeners from the similarity ratings are shown in figures 4.9 and 4.10, respectively.

	pno	tpt	hrn	tbn	cnt	flt	sax	vln	clo	vla
pno	0.6407	0.0156	0.037	0.0248	0.0709	0.0269	0.0809	0.0193	0.0548	0.0292
tpt	0.0059	0.2085	0.1006	0.1476	0.0562	0.0692	0.0574	0.0729	0.1386	0.143
hrn	0.0112	0.0895	0.1919	0.1448	0.1053	0.0911	0.1119	0.0773	0.0865	0.0905
tbn	0.0063	0.123	0.1248	0.1765	0.0696	0.0838	0.0686	0.0911	0.1321	0.1241
cnt	0.0347	0.0597	0.1123	0.0978	0.2233	0.1127	0.1567	0.0753	0.0671	0.0605
flt	0.011	0.075	0.0919	0.1029	0.1099	0.2261	0.0848	0.1016	0.092	0.1048
sax	0.0452	0.0585	0.0987	0.079	0.1744	0.0958	0.2455	0.0662	0.0685	0.0681
vln	0.0074	0.0672	0.0947	0.1143	0.0714	0.1033	0.0785	0.2464	0.1245	0.0923
clo	0.0073	0.1243	0.0799	0.1392	0.0548	0.0741	0.0487	0.1121	0.2001	0.1595
vla	0.008	0.1393	0.0761	0.1309	0.0604	0.0929	0.0528	0.082	0.1612	0.1965

Figure 4.9.
Average measured confusion matrix for NH listeners as a result of timbre similarity judgements.

	pno	tpt	hrn	tbn	cnt	flt	sax	vln	clo	vla
pno	0.515	0.0258	0.0616	0.0602	0.0884	0.0348	0.1351	0.017	0.028	0.034
tpt	0.0218	0.3181	0.1032	0.0874	0.0601	0.0572	0.0659	0.0641	0.1061	0.1161
hrn	0.0269	0.0636	0.2119	0.1488	0.1054	0.0643	0.1133	0.0765	0.081	0.1083
tbn	0.0172	0.0595	0.1584	0.2099	0.1192	0.057	0.081	0.0651	0.12	0.1126
cnt	0.0403	0.0481	0.127	0.1213	0.2426	0.0764	0.147	0.0584	0.067	0.0718
flt	0.016	0.0498	0.0868	0.1013	0.0944	0.2856	0.0928	0.0775	0.096	0.0999
sax	0.0683	0.0484	0.1351	0.095	0.1296	0.0876	0.2611	0.0513	0.0585	0.0652
vln	0.0159	0.0554	0.1009	0.1053	0.0777	0.076	0.0756	0.254	0.0972	0.1421
clo	0.0116	0.0639	0.0974	0.1402	0.0675	0.09	0.0565	0.0803	0.224	0.1687
vla	0.0125	0.0695	0.1231	0.1309	0.0752	0.0746	0.0507	0.1111	0.1424	0.2101

Figure 4.10.
Average measured confusion matrix for CI listeners as a result of timbre similarity judgements.

4.4 DISCUSSION

4.4.1 Measured timbre features

The results of the JND values found for the timbre perception features B, LRT and IRR for both NH and CI listeners in the first experimental study (section 4.3.1) generally agree with those of the literature on timbre perception abilities of CI users, which were found to be poor when compared to those of NH listeners (Gfeller *et al.*, 2002c; McDermott and Looi, 2004). This is illustrated by the substantially higher mean JND values found for CI listeners than for NH listeners. However, large SDs, particularly for the B JND values for CI subjects, can be noted.

Average JND values for CI listeners were more than four times those of NH listeners, which suggests that CI listeners seemed to have only approximately 25 % of the ability of NH listeners to perceive the features underlying timbre perception. This shows even poorer results than timbre identification studies, in which CI listeners were found to have approximately 50 % of the ability of NH listeners to correctly identify timbres (Gfeller *et al.*, 2002c; McDermott and Looi, 2004). The results of this study may be worse for CI listeners as a result of two factors: 1) the stimuli were simplified sounds that were synthesised according to a minimal number of parameters, thus restricting the number of cues that may otherwise be transmitted through real-world sounds to facilitate music perception, and 2) the definition of timbre perception in this study is encompassed by three specific features only, where in other studies, abilities of timbre perception as a whole have been reported. Additionally, it can be argued that timbre identification is not a direct measure of the perception of timbre, as music memory may affect the outcome of such studies. This introduces difficulties in directly comparing the overall results of this study to existing timbre perception studies for CI listeners.

4.4.1.1 Measured temporal timbre information

Rhythmic elements of music have been shown to be perceived better by CI listeners than melodic or pitch cues (Gfeller and Lansing, 1991). This suggests that temporal information is transmitted better than spectral information through a CI, which has also been shown in speech perception studies (McKay, 2005). Timbre perception studies by Gfeller *et al.* (1998),

Gfeller *et al.* (2002a) and Gfeller *et al.* (2002c) indicate that CIs are better at identifying percussion instruments, for example, the piano, than woodwind or brass instruments. This implies that the distinctive attack or rise time associated with percussion instruments serves as a valuable temporal cue in CI-mediated perception of music instrument timbre. Thus, as would be expected, the ability of CI listeners to discriminate the temporal feature of timbre, LRT, yielded JNDs most comparable to those of NH listeners.

Additionally, the JND of the temporal feature, LRT, has the lowest SD out of the three JND findings for both CI and NH subjects, indicating consistent results among CI listeners for LRT JNDs. This illustrates that for CI listeners, LRT is the most readily perceived of the three important timbre features, whereas B and IRR, which are based on spectral properties of the sound, appear to be transmitted less effectively. This corresponds to the findings of Kong *et al.* (2004), which showed that CIs currently provide enough spectral cues for speech perception in quiet, but are not adequate for music perception.

A more thorough interpretation of the comparable LRT findings for NH and CI listeners can be assisted by studies of temporal resolution tasks, including gap detection (Shannon, 1989) and amplitude modulation detection (Busby, Tong and Clark, 1993). Detection of gaps between sounds with JND values varying from 3 ms to 10 ms have been recorded in NH listeners (Clark, 2003), with the discrimination of gap duration JNDs comparable at values of 7 ms (Lister, Koehnke and Besing, 2000). Similar results were found for CI listeners, with gap detection and discrimination JNDs of 2 to 17 ms being recorded (Clark, 2003). The present study yielded a mean JND for the temporal feature LRT of 0.42 for NH listeners and a mean JND of 1.1 for cochlear implantees, corresponding to JND detections in the rise time of the sounds of approximately 2.63 ms for NH listeners and 12.55 ms for CI listeners, on average. Individual results for LRT for NH listeners range from 0.19 (1.56 ms) to 0.71 (5.18 ms), whereas for the CI listener group the LRT values range from 0.85 (7.1 ms) to 1.27 (18.8 ms). These LRT JND values compare well to the gap detection and discrimination JNDs reported by Clark (2003) and Lister *et al.* (2000).

4.4.1.2 Measured spectral timbre information

The spectrally associated timbre features (B and IRR) are conveyed substantially less effectively to CI listeners than to NH listeners, and with lower efficiency than LRT. This agrees with the measured results of Gfeller *et al.* (2002c) and McDermott and Looi (2004), in terms

of CI performance compared to NH performance for rhythm and pitch perception, as well as those of McKay (2005), in which the transmission of temporal information was found to be better than that of spectral information in CI listeners.

In the experimental tasks, two different types of spectral information were available to listeners: namely, B and IRR, relating to the global shape of the spectrum and the local shape of the spectrum, respectively. The large variations in B JNDs for CI listeners compared to NH subjects confirm that global spectral information is not generally transmitted as well to CI listeners as to NH listeners, and that the spectral information perceived is highly subject dependent. This may have been expected as a result of the individual differences in anatomical structure of the cochlea for each subject, the placement of the electrode array, and nerve survival in the cochlea.

Henry and Turner (2003) investigated the differences in spectral shape perception abilities of NH and CI listeners when listening to the same number of channels. Their results can be compared to the findings of this study for B, which is calculated from the general spectral shape of the sound. Henry and Turner (2003) found spectral shape perception for CI listeners to be poorer than for NH listeners, with average spectral component spacing JNDs for NH and CI listeners being around 400 Hz and 3000 Hz respectively, with large variations in JNDs for CI listeners, ranging from 800 Hz to 8000 Hz. This is comparable to the trend of the results found for B, where the CI listener group had a substantially larger JND value than the NH group (expressing B in Hz by multiplying by the fundamental frequency of 262 Hz gives us 124.23 Hz and 719 Hz, respectively), with a large SD in the JND values of B in the range of 64.66 Hz to 1373.88 Hz for CI listeners. The lower JND for NH listeners compared to that of the study of Henry and Turner (2003) may be explained by the fact that in their study, NH listeners listened to acoustic simulations of the sound, limited to 12 channels. The subject-specific nature of CI listener results, apparent from the large SDs of the the results, may explain the differences in JND values for B found in this study, compared to the results of Henry and Turner.

There is less consensus regarding the importance of IRR than the other two predominent features, B and LRT, for timbre perception. In the existing literature, the proposed third timbre feature is usually classified by one of two categories: spectro-temporal features, as in the case of spectral flux, or spectral features, as in the case of spectral fine structure or spectral spread, which is related to the shape of the spectrum (Caclin *et al.*, 2005). Although the study by Caclin *et al.* (2005) suggests that spectral irregularity is a more prominent dimension

than spectral flux, the authors conclude that it is nonetheless a less apparent dimension of timbre perception, possibly requiring further investigation. JNDs for IRR are not as severely affected in CIs as those for B, which can be attributed to IRR being extracted from the fine spectral structure of a sound, as opposed to the global spectrum of the sound in the case of B, making it a less salient perceptual timbre feature.

The present results for IRR can be compared to the results of a study by Henry and Turner (2003), in which the ability of implantees to resolve spectral ripples was also investigated and found to be significantly correlated with their ability to identify vowels. The results of Henry and Turner (2003) showed a spectral ripple resolution of approximately 1000 Hz for 50 % correct vowel identification, and a JND of up to 10000 Hz as the vowel identification dropped to 25 %. This poor resolution was illustrated by the larger JND obtained for IRR for CI listeners than for the NH group, where a resolution in the order of a few hundred Hz (for a fundamental of 262 Hz) would be required for IRR perception to be comparable to that of NH listeners. The main limiting factor preventing CI listeners from resolving frequencies is thought to be the differing degrees of excitation spread resulting from subject-specific neural survival patterns and pathological processes within the cochlea (McKay, 2005), which result in a blurring of spectral peaks or perceptual smearing in acoustic signals (Henry and Turner, 2003). Henry and Turner (2003) suggest that the ability to resolve spectral peaks may also be influenced by the compression of the acoustic dynamic range to the narrow electrical dynamic range in CIs.

A recent study by Emiroglu and Kollmeier (2008) attempted to quantify differences in object separation and timbre discrimination between NH and hearing-impaired listeners. The experiments determined JNDs of timbre in NH and hearing-impaired subjects along continua of “morphed” musical instruments with the attack times removed. JNDs of the morphing parameter which was investigated (also discussed in section 2.4.6.2) between pairs of sounds were found to be significantly lower for NH listeners than for severely hearing-impaired listeners. As discussed by Emiroglu and Kollmeier (2008), basilar membrane compression loss in sensorineural hearing-impaired listeners may lead to a distortion of mapping between the stimulus level that is presented and the stimulus level actually applied internally .This may make subtle intensity differences more audible, which could explain why the IRR results obtained for cochlear implantees are more comparable to those of NH listeners than in the case of the results for B. Slight changes in amplitudes of spectral components will not change the overall spectral envelope or value of B, whereas the spectral irregularity, IRR, may be substantially affected.

A study by Turner and Holte (1987), in which discrimination of spectral shapes in speech-like sounds was investigated for NH and CI listeners, may explain the necessary requirements for successful transmission of features B and IRR . Under certain conditions, higher, more prominent spectral peaks were required for CI listeners to perform equivalently to NH listeners. For those that did not achieve normal discrimination results at any level of increased spectral peak presentation, high frequency amplification of the stimuli (high-pass amplification) returned the JNDs to the NH range. For IRR, alternating frequency bands may need intensity amplification to make the spectral differences in adjacent harmonics more prominent to CI listeners.

Hopkins and Moore (2007) and Moore, Glasberg and Hopkins (2006) performed frequency discrimination studies and found that the ability of CIs to use temporal fine structure is poor. It was found that harmonics above the 5th were not resolved (Moore *et al.*, 2006). This may explain the poorer ability of CI listeners to perceive B and IRR than NH listeners, as only the first five spectral components would likely be used for perceptual judgements. Moore *et al.* (2006) also found a reduced ability of CI listeners to use temporal envelope cues, which could also explain the poorer abilities of CI listeners to perceive LRT, when compared to NH listeners.

The fundamental frequency, which relates to features B and IRR, is an important feature in the perception of musical instruments sounds, with the central auditory processing thereof facilitated by either spectral or temporal methods (McKay, 2005). The difficulties that CI listeners experience in perceiving timbre can primarily be explained by the restrictions of existing processing strategies (SPEAK, ACE or CIS), in which the fixed overlapping filter bands that are used limit the number of harmonic components that can be resolved and the identification of the harmonic components. Additionally, phase shifts between electrodes positioned close together may lead to incorrect perception of the fundamental frequency. Even with the provision of more perceptual channels in CIs, temporal information would probably also have to be conveyed by the analysis channels at the correct tonotopic place in the cochlea, as discussed by McKay (2005). Additionally, factors such as the smoothing of filter outputs and the uncontrolled phase differences that occur on electrodes placed nearby (McKay and McDermott, 1996) need to be addressed to better convey the fundamental frequency and harmonic components of a sound.

4.4.2 Timbre similarity ratings

The results of the second experimental study, in which timbre similarity judgements for both NH and CI listeners were investigated, are shown in section 4.3.2. These results can be compared to some studies reported in the literature: the confusion matrices of figures 4.9 and 4.10 can be compared to previous timbre perception results for NH and CI listeners.

As discussed by Donnelly and Limb (2009), NH listeners regularly mistake musical instruments from the same family, which can be seen from the higher confusions, found in this study and shown in figure 4.9, between the string instrument family consisting of the violin, cello and viola and between instruments of the brass family, specifically between the French horn and trombone. Higher confusions were also noted between instruments of the woodwind family, particularly between the clarinet and flute. Similar findings were reported by Gfeller *et al.* (2002c), where the highest confusions in the identification of instruments for NH listeners were found between the woodwind instruments, the clarinet and saxophone, and between the string instruments, the violin and cello. Even though the NH listeners were exposed to processed instrument sounds, the listeners appear to employ the same timbre perceptual cues that would be used in listening to unprocessed sounds to rate the similarity of two sounds.

The experimental confusion matrix of the CI listeners as shown in figure 4.10 shows more scattered error patterns not necessarily corresponding to instrument family, as was also found by Donnelly and Limb (2009). Although high probabilities of confusions occur between instruments of the brass family, such as the French horn and trombone, as well as between string family instruments, high confusions are also found between instrument families, for example, between the clarinet and trombone and the cello and trombone. Instrument sound confusion matrices obtained by Gfeller *et al.* (2002c) showed the largest confusions for CI listeners between string instruments, corresponding well to the experimental confusions found for CIs in this study. Additionally, Limb (2006) reported percussion instruments to be the most readily identified by CI listeners, corresponding to the generally low confusions of the piano with other instruments found in this study, as illustrated in figure 4.10. Hall and Beauchamp (2009) discuss the role of the rise time of sounds, where instruments with very abrupt rise times, such as the piano, may serve as a perceptual reference against which all other stimuli are evaluated. This agrees with the results of this study, in which the piano was always found to be the most distinct from the other instruments.

In addition to the distinguishing temporal properties of the piano that produces less confusion with other instruments, the normalisation of the peaks of the instrument sounds should also be noted as potentially enhancing dissimilarities between the piano and other instrument sounds, and indeed amongst many of the musical instrument timbres. Although peak normalisation of the sounds assisted in balancing the perceived loudness of the different musical instrument sounds, this does not completely balance the sounds in terms of loudness, as the energies of the sounds differ. This indicates that perceptual differences in the loudness of the sounds may still be present and that this could have affected the results of the similarity ratings, as any difference in loudness would imply a difference between two sounds.

Although the methodology of the timbre perception experiments of this study differed from the timbre recognition tasks reported in literature, the results showed similar trends to previous timbre perception findings, as discussed above. The experimental findings of this study could thus be used as data with which to compare the predictions of the developed model of timbre perception.

4.5 CHAPTER SUMMARY

The experimental component of this study was presented in this chapter, consisting of two experiments performed with both NH and CI listeners. The discrimination task results where the JNDs of timbre features B, LRT and IRR were found for each listener were presented. These were used in obtaining predictions from the model of timbre perception, as discussed in chapter 5. The results of the similarity judgements of musical instrument sounds were also provided. Chapter 5 presents the analysis and comparison of these results to the predictions of the timbre perception model. This chapter discusses the outcomes of the experimental components of this study, which provide an entry point into achieving a quantitative understanding of the timbre perception abilities of CI listeners, by providing measurable results in terms of timbre perception features. These measurable results were then used in the development of a model of timbre perception, as discussed in chapter 5.

CHAPTER 5

MODELLING OF TIMBRE PERCEPTION

A large part of the work described in this chapter was presented at the CI 2010 conference (Hanekom and Hugo, 2010). The work in this chapter will also be submitted to Ear and Hearing in the form of a journal article as a continuation of the work discussed in chapter 4.

5.1 CHAPTER OBJECTIVES

The experimental component of this study was presented in chapter 4. The findings of the first experiment (section 4.3.1) were used in the development of the model of timbre perception. The model predictions could then be compared to the the findings of the second experiment (section 4.3.2). This chapter presents the implementation of the model of timbre perception, which was performed in Matlab version 2007b, as well as the results of the model predictions compared to the experimental results. An explanation of the analysis techniques used to interpret the results and allow for comparisons to be made between predicted and measured results for NH and CI listeners, to validate the measurements and models of this study are also presented in this chapter. A detailed discussion is provided regarding the outcomes of the model of timbre perception, along with considerations made in the development of the model. Comparisons of the results of the study with existing literature will be explored and implications of the timbre perception model outcomes will also be presented.

5.2 METHODS

5.2.1 Timbre perception model

A model of timbre perception was developed in an attempt to adequately predict the results of timbre perception experiments, and, in this study specifically, to predict the outcome of timbre similarity rating experiments as described in section 4.2.2. The model was based on the three important timbre features, B, LRT and IRR, and the definitions of both the original and processed sounds by these three features. The model was structured on the premise that the features B, LRT and IRR formed a three-dimensional orthogonal space, or timbre space[1]. The first step in implementing the model was to extract the features from each of the original and processed sounds, using the methods described in section 3.4. As musical instrument timbres are generally recognised by the information that is transmitted by B, LRT and IRR, it is implied that when different musical timbres have similar B, LRT or IRR values, that there is a possibility that these timbres may become confused with each other.

In the case of predicting confusions amongst vowels, three-dimensional Euclidean distances may be used, as illustrated by Conning (2005). The Euclidean distances are measured between all the vowels in the vowel space to predict the amount of confusion between each pair. However, for this study, a more detailed approach, based on the work of Svirsky (2000), is used to predict confusions between instrument timbres, by expanding on the use of Euclidean distances alone. For predictions of confusions between vowels, a normalisation of the vowel space, generally defined by the duration of the vowel and the first two formant frequencies, F_1 and F_2, is usually performed before the Euclidean distances are calculated (van Wieringen and Wouters, 1999). Lobanov's z-score transformation (Adank, Smits and Van Hout, 2004) is a possible choice of normalisation that may be implemented, as it allows for the comparison of vowels across various conditions. The processing of sounds may cause an offset to be added to a specific vowel space, which is removed by normalisation. However, Lobanov's z-score transformation is a vowel-extrinsic procedure, and so cannot be used to normalise the timbre space. Additionally, no normalisation procedures specific to timbre spaces have been developed. As a result, simply calculating the Euclidean distances between instrument sounds in the timbre space would be questionable, as the dimensions are not normalised and cannot necessarily be measured relative to one another.

[1] a timbre space is defined as a multidimensional space where a number of instrument sounds are plotted as a function of their signal characteristics, defined by B, LRT and IRR.

Therefore, using the model implemented by van Zyl (2008) and based on the model by Svirsky (2000), a volume surrounding each of the points that represent a sound in the timbre space can be applied to calculate the probabilities of confusing each sound with another. The volume surrounding each sound in the timbre space is modelled by an ellipsoid, with the dimensions of each of the three axes thereof corresponding to the JND values found for each of the timbre space dimensions B, LRT and IRR, from the experimental study in section 4.2.1. A basic illustration of the structure of the model of timbre perception is given in figure 5.1, with an arbitrary musical instrument sound in the timbre space, in terms of B, LRT and IRR, represented by the black circle. An ellipsoid is constructed around the position of the instrument sound in the three-dimensional space, using the JND values obtained for each of the timbre features as the dimensions of the ellipsoid in each corresponding timbre dimension. Such ellipsoids are constructed around each musical instrument sound, with the amount of overlap between two ellipsoids indicative of the amount of confusion that is likely to occur between two specific musical instrument sounds.

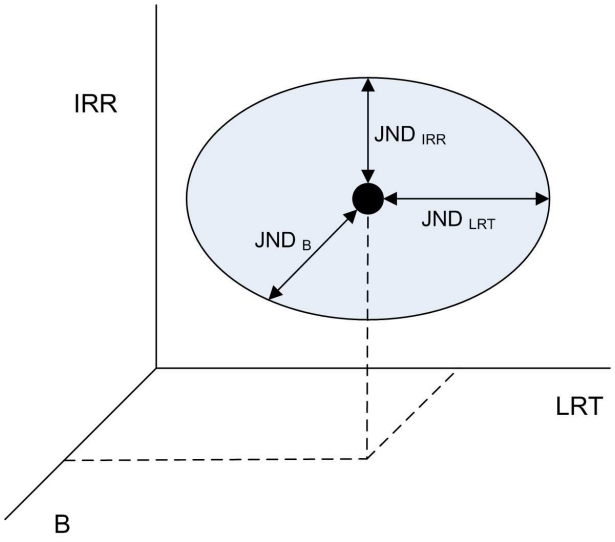

Figure 5.1.
Illustration of the structure of the model of timbre perception. For each musical sound represented in a three-dimensional timbre space, an ellipsoid is constructed, using the JND values obtained for each of the timbre features as the dimensions. The units for each of the axes for B, LRT and IRR are defined in sections 3.4.1 to 3.4.3

As an existing model was used to predict the probabilities of confusions between the instrument sounds, a gain factor of 1 for the JND values was the only model parameter that was set in the development of the model of timbre perception. All other parameters were inherent in the model by van Zyl (2008).

To calculate the probability of confusions from the overlap of the ellipsoids constructed around each sound, signal detection theory (Gelfand, 1990; Green and Swets, 1966) was utilised. This method is commonly applied to psychophysics, and was thus suitable for the purposes of this study. The basic idea of the model was to assign a probability density function (pdf) to each point or sound in the timbre space of figure 5.7, to create a volume around each instrument sound that more accurately represents the space in which the sound would be confused with any of the other sounds in the timbre space. A four-dimensional pdf is thus necessary to represent each sound in the timbre space (to represent the three variables of the timbre space) and all the pdfs were generated to have a Gaussian distribution. In this study, the mean of the pdf was obtained from the three-dimensional coordinates of the instrument sound in the timbre space, represented by B, LRT and IRR. The variances of the pdfs were calculated from the JNDs obtained from the first experimental study in section 4.2.1, as opposed to using uncertainty factors calculated from a processing component in the method of van Zyl (2008). Once the pdfs were constructed for each sound, the amount of confusion between two sounds could be predicted.

The model applied in this study used three variables or dimensions (B, LRT and IRR), thereby creating a four-dimensional pdf for each sound. However, for the purpose of illustration, an explanation of the model implementation will be given for a one-dimensional variable, resulting in a two-dimensional pdf. The Gaussian distribution, given by $f(x)$ in equation 5.1 for one dimension is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \tag{5.1}$$

where μ is the mean of the distribution, and σ is the variance. A pdf was calculated for each of the instrument sounds. The probability value for each element in a confusion matrix; that is, the probability of the listener giving a specific response given all possible responses, was calculated individually. This was achieved by integrating the pdf of the stimulus (or particular instrument sound) from a certain decision point, as shown in figure 5.2. The decision point is chosen as the point of intersection between the stimulus pdf and the possible response (a different instrument sound in the space that the stimulus may be confused with) that is being calculated.

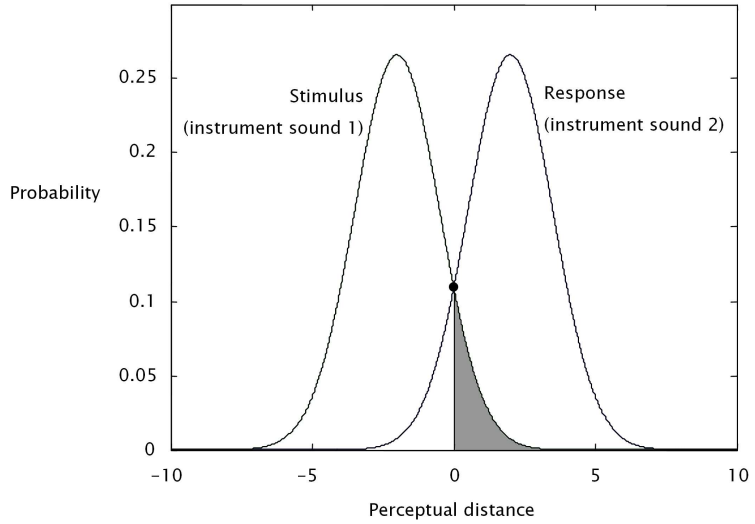

Figure 5.2.
Illustration of two-dimensional probability density functions representing the stimulus (the instrument sound investigated) and response (a different instrument sound with which the one under investigation may be confused). The probability of confusing these two sounds or giving the incorrect response (shaded area) is in this case calculated from 0 to the end of the stimulus pdf.

The more overlap there is between the pdfs of two sounds, the greater the probability that the listener will confuse the two sounds. The distance between the centre points of the pdfs is found from the Euclidean distance in terms of the timbre features B, LRT and IRR in the timbre space. Figure 5.2 only shows the case when one variable is used, as opposed to the three used in this study. Thus, the listener must integrate these various components into a single decision. To model this, trivariate random variables are used instead of the univariate random variable as is the case in figure 5.2, so that an observation is a point in three-dimensional space as opposed to a point on a line. The Gaussian distributions can thus be expanded to a matrix form for the trivariate random variable calculations. van Zyl (2008) provides the detailed equations for these calculations, as well as those for finding the single decision point from the trivariate pdfs.

A visual representation of the trivariate random variable calculations and resulting pdfs would require a four-dimensional illustration. Therefore, the ellipsoid representations are used to visualise the distances between the instrument sounds and the possible confusions between these instrument sounds. Figure 5.3 illustrates the ellipsoid representation for three arbitrary musical instrument sounds in the timbre space. In the proposed model, the closer the ellipsoids are that surround each of the sounds, the larger the probability that the sounds will be confused with each other. If two ellipsoids do not intersect, there is a smaller probability that the instrument sounds would be confused. Thus, from figure 5.3, instrument sounds 2 and 3 will be more likely to be confused with each other than either of these two sounds would be with instrument sound 1. The pdf calculations can then be applied to determine the probabilities of confusion predicted between each instrument sound pair.

In developing the model of timbre perception, a number of assumptions were made which were likely to impact the model predictions. Firstly, the JND values were only calculated for one synthesised reference tone, and therefore the same JND values were used to create identical ellipsoids around each of the musical instrument timbres. In addition to this, the ellipsoids were symmetrical, a feature that is also inherent in the model by van Zyl (2008), and thus the assumption that the JNDs are the same on either side of a sound in a particular dimension in the timbre space had to be made.

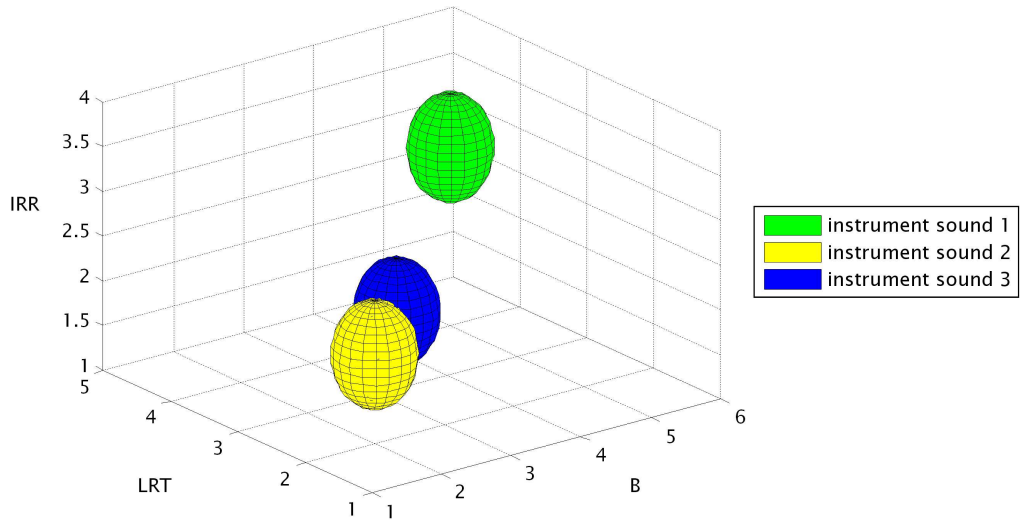

Figure 5.3.
A three-dimensional timbre space generated by the model of timbre perception for three arbitrary musical instrument sounds represented by ellipsoids, with units for the axes B, LRT and IRR as defined in sections 3.4.1 to 3.4.3.

A full description of the method used to calculate the predicted confusion matrices is given in van Zyl (2008), based on the model developed by Svirsky (2000). The results of the predicted confusions obtained from the model of timbre perception, as well as the resulting JNDs for B, LRT and IRR, are shown in section 5.3.1.

5.2.2 Analysis techniques

The analyses of the predicted and experimental results were based on methods applied to speech perception research in CIs for vowels and consonents. Details of the analyses of the results are given in section 5.3.2, and a brief overview of the methods utilised are discussed in the paragraphs that follow.

Relative information transmission scores are a common method of analysing stimulus-response results of psychoacoustic experiments, and have been used extensively in research for speech perception in CI listeners. The general procedure involves entering responses to stimuli into

confusion matrices, which are stimulus-response matrices that indicate which phonemes have been confused with which. The confusion matrix contains information about which cues have been transmitted to the auditory system and which have been masked. Confusion matrices are analysed with feature information transmission analysis (FITA) techniques. It is well known that for the identification of vowels, the first two formants and the duration of the vowel are necessary cues (van Wieringen and Wouters, 1999). For the identification of consonants, both acoustic and articulation properties are necessary cues. Acoustic properties include envelope variation of the consonant (van Tasell, Soli, Kirby and Widin, 1987), the ratio of the minimum to peak energy of the consonant and the duration of the consonant (van Wieringen and Wouters, 1999). Articulatory classification of consonants includes the following categories: plosive or non-plosive, voiced or voiceless, place of articulation (front, middle or back of the mouth), nasal or non-nasal, liquid or non-liquid, and fricative or non-fricative (Miller and Nicely, 1955; Wang and Bilger, 1973).

For analysis purposes, each phoneme is first classified into one of several categories for each cue. After classification, each cue is looked at separately. The confusion matrix is collapsed into the number of categories available for that cue. For example, for place of articulation there are three categories: 1 = front, 2 = middle and 3 = back. The relative information transmitted through each cue can then be calculated by the ratio of the transmitted information calculated from the confusion matrix to the maximal possible information transferred by the stimuli and categories under test.

Using a similar approach, confusion matrices for timbre perception may be constructed and analysed to indicate the amount of information transmitted through each cue important for timbre perception. As there are no existing results for confusions between timbres for CI listeners where important cues have been extracted, the important cues assumed will be the same as those for NH listeners, namely B, LRT and IRR. FITA results obtained from similarity judgements of musical timbres by CI listeners can then show which timbre features or cues convey the most information, and how well they convey information about timbre to CI listeners.

5.3 RESULTS

5.3.1 Results of the timbre perception model

The timbre features B, LRT and IRR were extracted for both the original and processed versions of the 10 musical instrument sounds as a basis on which to formulate the model of timbre perception. Table 5.1 shows the extracted timbre features B, LRT and IRR for the sounds.

Table 5.1.
Values for B, LRT and IRR extracted for the original and processed musical instrument sounds, with units for B, LRT and IRR defined as partial index, log(s) and log(dB), respectively.

		Original musical timbres			**Processed musical timbres**		
		B	LRT	IRR	B	LRT	IRR
Piano	**pno**	2.6523	1.8027	1.5840	1.8097	1.2093	1.2878
Trumpet	**tpt**	5.8998	2.8710	1.2270	2.2031	2.8470	1.5759
French horn	**hrn**	4.3149	2.5357	1.7323	1.8638	2.8147	1.2155
Trombone	**tbn**	3.4961	2.6073	1.0274	1.9855	2.6056	1.2535
Clarinet	**cnt**	3.0455	2.9564	1.9858	1.7204	3.1122	1.2485
Flute	**flt**	4.0928	2.9285	1.5072	1.9199	2.8027	1.2802
Saxophone	**sax**	3.8301	1.7058	1.7850	1.6182	1.9788	1.4433
Violin	**vln**	5.9806	3.0502	2.1186	1.8960	3.1625	1.1597
Cello	**clo**	4.5775	2.8361	1.6376	2.1106	2.7855	1.3037
Viola	**vla**	4.5712	3.0988	1.9179	2.1409	3.0883	1.4015

The original and processed sounds are illustrated graphically in two dimensions for combinations of the three timbre features in figures 5.4, 5.5 and 5.6. It can be noted that processed sounds are grouped closer together than the original sounds, showing that the probability of confusions between sounds is increased substantially with processing through a CI.

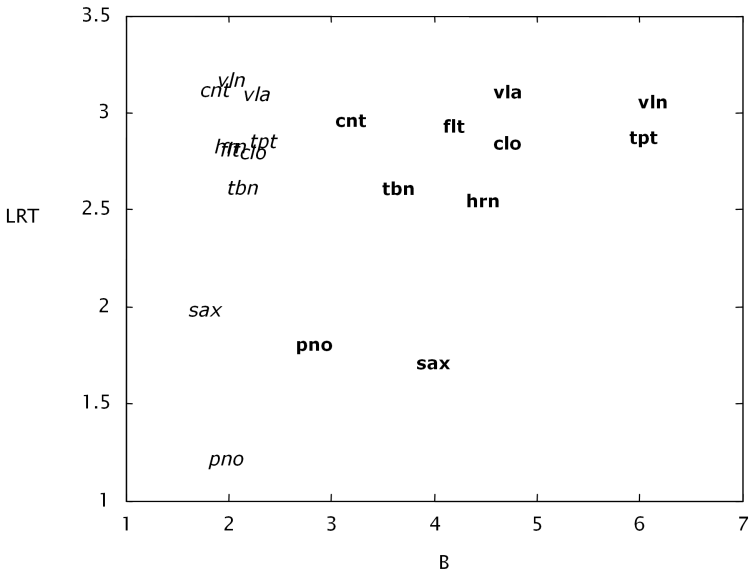

Figure 5.4.
Representation of the original instrument sounds (bold) and processed instrument sounds (italics) represented by timbre dimensions B and LRT, with units for the timbre features as given in table 5.1.

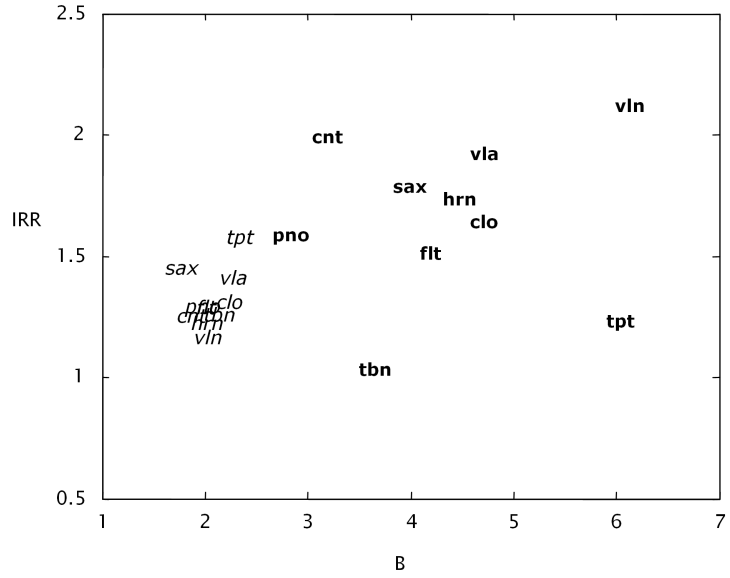

Figure 5.5.
Representation of the original instrument sounds (bold) and processed instrument sounds (italics) represented by timbre dimensions B and IRR, with units for the timbre features as given in table 5.1.

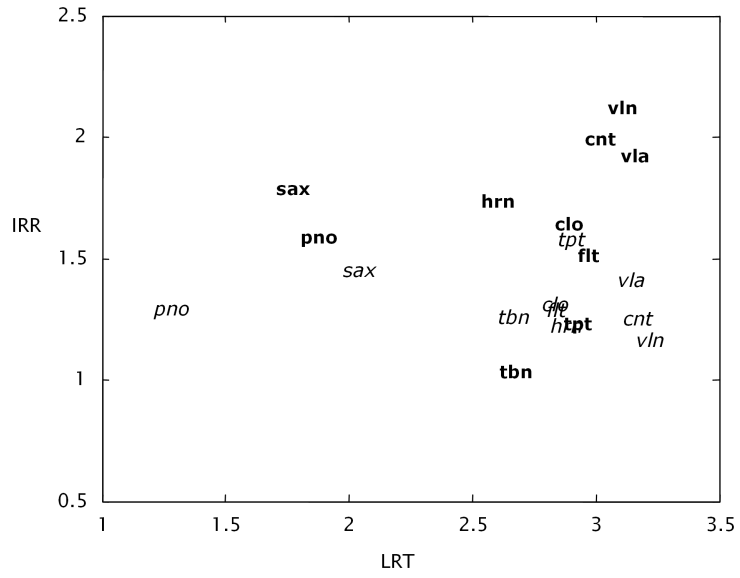

Figure 5.6.
Representation of the original instrument sounds (bold) and processed instrument sounds (italics) represented by timbre dimensions LRT and IRR, with units for the timbre features as given in table 5.1.

Figure 5.7 graphically represents the values of table 5.1 in three dimensions by displaying the original and processed musical instrument sounds in terms of B, LRT and IRR.

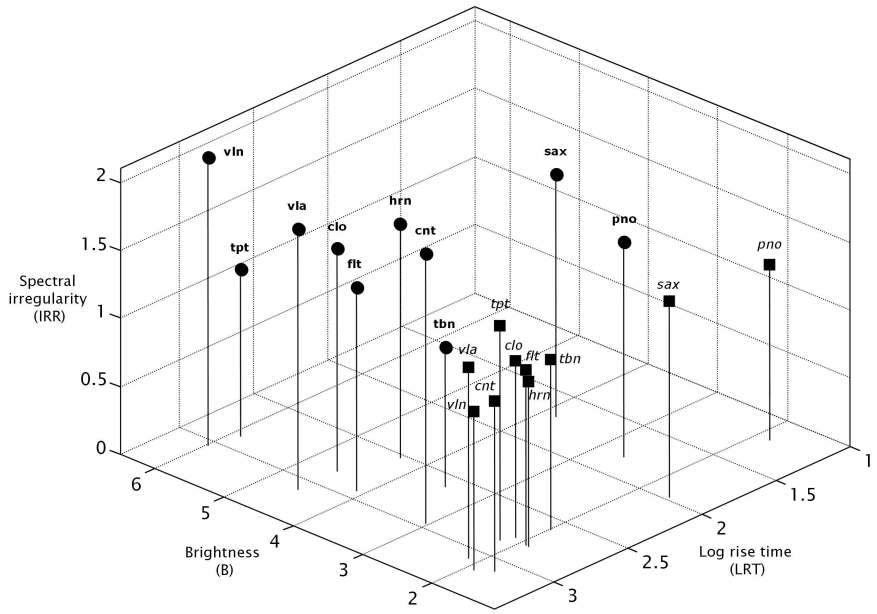

Figure 5.7.
Timbre space of the original (circle) and processed (square) instrument sounds represented by timbre dimensions B, LRT and IRR, with units for the timbre features as given in table 5.1.

The visual representation of the original and processed sounds in figure 5.7 provides an indication of which instrument sounds would be more likely to be confused with one another. Sounds that lie closer together would have a higher probability of being confused, while those that lie further apart would be less likely to be confused with each other. Predictions about which music timbres will be confused can be made visually by inspecting the timbre space. From this, a physical measure can be used to quantify these predictions.

In all of the two-dimensional representations for the sounds processed by the acoustic model, the piano is the most distinct sound, as it is most distant from the others. This can also be seen from the experimental confusion matrices obtained for both NH and CI listeners, with low probabilities of confusion for the piano compared to the other sounds. In both the NH and CI experimental results, the saxophone was found to be the most likely to be confused with the piano. This can be explained by the two-dimensional timbre spaces of figures 5.4

and 5.6, as the processed saxophone sound is also distinct from the other instruments, but generally lies closer to the piano than to the other instrument sounds in the timbre space.

The experimental confusions for NH listeners (figure 4.9) and CI listeners (figure 4.10) show large confusions between brass instruments, specifically between the trombone and French horn, with slightly lower confusions for the trumpet. This would be expected, as shown in the two-dimensional timbre space representations of the processed sounds, in which in figures 5.4 and 5.6 in particular the processed brass instruments are grouped in the same vicinity, but the trumpet is always situated slightly separately from the French horn and trombone.

For both NH and CI listeners, high confusions were found among the family of string instruments (figures 4.9 and 4.10), specifically between the viola and the cello and the violin and the cello, with the confusion between the violin and viola being somewhat lower. The two-dimensional representations on which the predictions are based (figures 5.4 to 5.6) illustrate these experimental results by showing viola-cello and violin-cello combinations to be located in close proximity, but with the viola and violin found to be further apart.

For NH listeners, dominant confusions are also present between woodwind instruments, namely between the saxophone and clarinet and the flute and clarinet. CI listeners also found the saxophone and clarinet very similar. These confusions are not illustrated as clearly by the two-dimensional representations, especially for the saxophone. Substantial confusions were also found between the French horn and clarinet for both NH and CI listeners, which is a confusion between families of instruments, and this is explained and illustrated by the close proximity of these processed instruments in the two-dimensional timbre spaces of figures 5.4 to 5.6. High incidences of confusions between the clarinet and trombone were found for CI listeners, and can only be explained by figure 5.5. Additionally, the clarinet and French horn were also found to be very similar for both NH and CI listeners, again due to the similar location of these processed sounds in the timbre space in figures 5.4 to 5.6.

An additional factor that may have affected the grouping of the sounds in the timbre space could have been the exclusion of signal pre-emphasis in the developed acoustic model. Without signal pre-emphasis, the lower frequency channels would have been more likely to be selected and included in the stimulation pattern. This could explain the tight grouping of the processed instrument sounds in the B and IRR dimensions in figures 5.4 to 5.6, as the frequency channels selected by the peak-picking method of the ACE strategy for each sound would have been in a similar lower range.

The model for predicting timbre perception was not based solely on the distances between the sounds in the timbre space, but this formed a basis for the model structure. A brief representation of the timbre space of the sounds compared to the confusion matrices obtained from both the NH and CI experimental data already shows clear similarities, substantiating the choice of the timbre space as the model foundation.

Predictions of timbre perception were made using the JND results for the timbre features as shown in figure 4.7. The individual JND values of each subject are used as dimensions of the ellipsoid centred around each point representing an instrument sound in the timbre space. A predicted confusion matrix for each subject can thus be obtained by the method of van Zyl (2008), as discussed in section 5.2.1. Figure 5.8 constitutes a visual representation of the probability of confusions between each of the musical instrument timbres for both the unprocessed and processed sounds. These ellipsoids are calculated using the average JND values obtained for each timbre feature for NH and CI listeners, to illustrate the differences between the processed and unprocessed sounds in the timbre space, as well as the differences in JNDs for the NH and CI subjects. Ellipsoids calculated in this way were used to predict the confusion matrices that are obtained from the similarity ratings data.

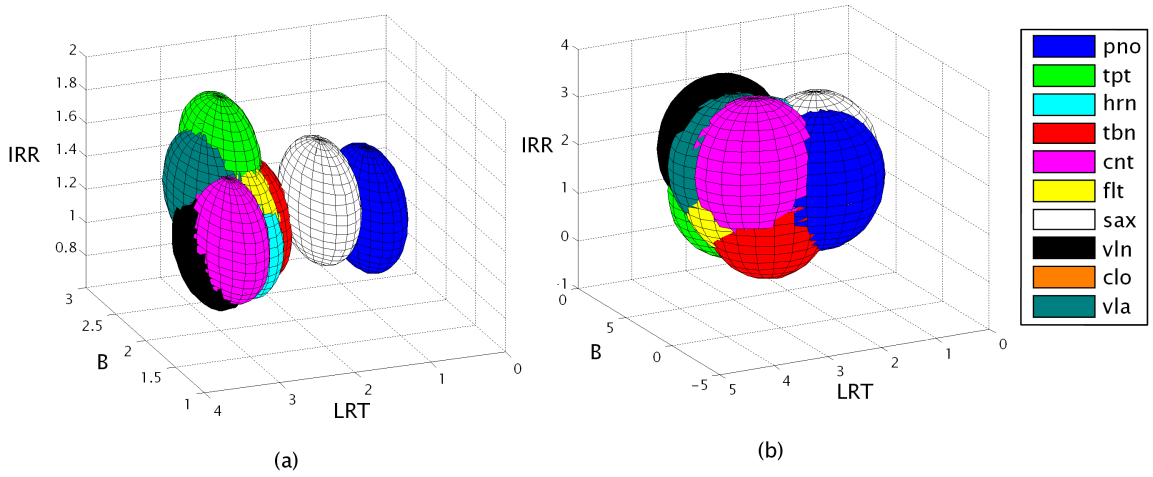

Figure 5.8.
A three-dimensional timbre space generated by the prediction model for (a) NH and (b) CI listeners. The predictions for NH listeners were modelled on the timbre dimensions for the instrument sounds processed through the acoustic model and the predictions for CI listeners were modelled on the timbre dimensions for the original instrument sounds. Units for the axes of the timbre features B, LRT and IRR are as given in table 5.1.

Individual JND values for B, LRT and IRR for each listener were used to obtain subject-specific predicted confusion matrices, in order to compare these with individual experimental confusion matrices obtained for each subject from the similarity ratings. Each row of the confusion matrix is normalised with respect to the sum of that row. Averages of both the NH subjects' and CI subjects' predicted confusion matrices were used in the analysis of the results. Figures 5.9 and 5.10 show the average predicted confusion matrices obtained from the model of timbre perception for NH and CI listeners respectively.

	pno	tpt	hrn	tbn	cnt	flt	sax	vln	clo	vla
pno	0.6582	0.0229	0.0262	0.0285	0.0296	0.0224	0.1308	0.0246	0.0339	0.023
tpt	0.0109	0.2619	0.0841	0.0927	0.0658	0.106	0.0362	0.0626	0.1353	0.1445
hrn	0.0103	0.0619	0.1861	0.129	0.1119	0.1599	0.0362	0.1018	0.1204	0.0825
tbn	0.0115	0.0777	0.1426	0.2065	0.0517	0.1583	0.0497	0.0786	0.1453	0.078
cnt	0.0131	0.0588	0.1371	0.0573	0.2331	0.1277	0.0239	0.166	0.0859	0.0971
flt	0.0086	0.0722	0.1472	0.1335	0.0987	0.1715	0.0362	0.0919	0.1529	0.0874
sax	0.085	0.0508	0.0639	0.0874	0.038	0.0674	0.4888	0.0305	0.0618	0.0265
vln	0.0105	0.0556	0.12	0.0838	0.1619	0.1146	0.0191	0.2293	0.0971	0.1082
clo	0.0129	0.0964	0.1199	0.1292	0.0695	0.1622	0.0354	0.0796	0.1837	0.1111
vla	0.0104	0.1242	0.0948	0.0844	0.0949	0.1118	0.0201	0.1059	0.1301	0.2233

Figure 5.9.
Average predicted confusion matrix for NH listeners, calculated using the model of timbre perception.

	pno	tpt	hrn	tbn	cnt	flt	sax	vln	clo	vla
pno	0.1872	0.0537	0.0989	0.113	0.1115	0.0941	0.1466	0.0462	0.0791	0.0697
tpt	0.0499	0.1825	0.1037	0.0734	0.0627	0.0995	0.0673	0.1295	0.1222	0.1093
hrn	0.0749	0.0795	0.1361	0.0964	0.0892	0.1151	0.1008	0.0759	0.1181	0.1141
tbn	0.0966	0.0674	0.1096	0.1559	0.1031	0.1177	0.0939	0.0587	0.1044	0.0925
cnt	0.0961	0.0568	0.1033	0.106	0.1605	0.1146	0.0909	0.0637	0.1	0.108
flt	0.0719	0.0782	0.116	0.104	0.0993	0.1372	0.0856	0.0717	0.1242	0.1118
sax	0.128	0.0635	0.1183	0.0967	0.0922	0.0996	0.1611	0.0573	0.1017	0.0816
vln	0.0446	0.1321	0.0991	0.0669	0.0707	0.0917	0.0629	0.1858	0.1147	0.1314
clo	0.0604	0.0946	0.1176	0.0919	0.087	0.1225	0.0862	0.0868	0.1353	0.1177
vla	0.0552	0.0871	0.1186	0.0831	0.0975	0.1149	0.0735	0.1053	0.1231	0.1416

Figure 5.10.
Average predicted confusion matrix for CI listeners, calculated using the model of timbre perception.

5.3.2 FITA results

FITA was performed on the confusion matrix results of the timbre similarity rating experiments, as well as on the predicted confusion matrices obtained from the model of timbre perception. A FITA approach allows the confusion matrices to be collapsed so that their similarity in each timbre dimension may be compared individually. FITA also provides information on the cues that have and have not been received, information that formed an important part of this study, in which the transmission of timbre features to NH and CI listeners for both predicted and measured data were investigated. The FITA procedure for calculating the amount of information transmitted is explained next. Taken from Miller and Nicely (1955), the mean logarithmic probability (MLP) equation, as shown in equation (5.2), is a measure of covariance between input and output. If the input variable is *x*, with probability p_i and *i* = 1,2,...,k, then the input is defined as

$$\mathrm{MLP}(x) = E(-\log p_i) = -\sum_i p_i \log p_i. \tag{5.2}$$

If the logarithm is taken to base 2, then the measure can be called the number of binary decisions needed on average to specify the input, or number of bits of information per stimulus. A similar expression is defined for the output *y*, with values *j* = 1,2,...,m. The number of decisions needed to specify the particular stimulus-response pair is MLP(*xy*), with p_{ij} being the probability of the joint occurrence of input *i* and output *j*.

A measure of covariance of input with output is given by equation 5.3 with *T(x;y)* referred to as the transmission from *x* to *y* in bits per stimulus.

$$T(x;y) = \mathrm{MLP}(x) + \mathrm{MLP}(y) - \mathrm{MLP}(xy) = -\sum_{ij} p_{ij} \log \frac{p_i p_j}{p_{ij}} \tag{5.3}$$

In practice, the true probabilities are not known and are estimated from the relative frequencies obtained experimentally, giving equation 5.4 from van Tasell *et al.* (1987) as

$$U = -\sum_{ij} \frac{n_{ij}}{n} \log_2 \frac{(n_i/n)(n_j/n)}{n_{ij}/n}, \tag{5.4}$$

with n being the total number of observations, n_i the frequency of the stimulus, n_j the frequency of the response and n_{ij} the frequency of the joint occurrence of a particular stimulus-response pair.

The maximum available information is given by equation (5.5) as

$$U_{max} = -\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n}, \tag{5.5}$$

giving the relative transmitted information as

$$U_{rel} = U/U_{max}. \tag{5.6}$$

If the response is closely correlated with a specific stimulus, U_{rel} will be close to unity as the specific feature will be transmitted well. The relative information transmitted is the ratio of the transmitted information calculated from the confusion matrix to the maximum possible amount of information transferred by the stimuli (instrument timbres in this study) and the features being tested (B, LRT and IRR) (van Wieringen and Wouters, 1999). In this study, the stimuli for the instrument sounds are classified as in tables 5.3 and 5.4 for the original and processed instrument sounds, respectively, and the response is a confusion matrix, either from the experimental study or predicted from the model. In this way, $T(x,y)$ can be calculated for each of the features given in tables 5.3 and 5.4. The classifications of the timbre features are determined using the ranges given in table 5.2. The assignment of the sounds to the different timbre feature ranges was made to encompass the range of features for both the unprocessed (table 5.3) and processed (table 5.4) sounds. The categories were selected to optimise the grouping of sounds from the same instrument families as far as possible for the unprocessed instrument sounds, in terms of B, LRT and IRR. The categories were chosen so that each of the original instrument sounds in each family of instruments fell within the same category for at least two of the three timbre features when compared to any other instrument sound in that same family. For example, for the original string instruments consisting of the violin, cello and viola, the B and IRR values of all three instrument sounds fell within the same category.

Table 5.2.
Ranges of B, LRT and IRR used for the classification of important timbre features into categories for FITA results to be obtained.

	B	**LRT**	**IRR**
1	< 1.9	< 1.9	< 1.25
2	1.9 - 3	1.9 - 2.9	1.25 - 1.6
3	3 - 4.5	> 2.9	> 1.6
4	> 4.5	-	-

Table 5.3.
Classification of important timbre features to be used in FITA for original instrument sounds.

	pno	**tpt**	**hrn**	**tbn**	**cnt**	**flt**	**sax**	**vln**	**clo**	**vla**
B	2	4	3	3	3	3	3	4	4	4
LRT	1	2	2	2	3	3	1	3	2	3
IRR	2	1	3	1	3	2	3	3	3	3

Table 5.4.
Classification of important timbre features to be used in FITA for instrument sounds processed through the acoustic model.

	pno	**tpt**	**hrn**	**tbn**	**cnt**	**flt**	**sax**	**vln**	**clo**	**vla**
B	1	2	1	2	1	2	1	1	2	2
LRT	1	2	2	2	3	2	2	3	2	3
IRR	2	2	1	2	1	2	2	1	2	2

The instrument sounds are grouped together according to their classifications, to determine the percentage information transmitted through a specific timbre characteristic. The confusion matrices can then be analysed using these classifications, to find how much information is conveyed through each timbre feature. This will also serve as basis for comparison between the model predictions and the experimental results.

5.3.2.1 FITA results of experimental data

FITA was performed on the measured results to determine how much information was transmitted to the listener through each timbre feature. This also allowed for comparisons to be made regarding which characteristics were transmitted most effectively to CI listeners and to NH listeners through the acoustic simulations. The average FITA results obtained for NH and CI listeners from the timbre similarity rating experiments are shown in figure 5.11, with means and SDs of the percentage information transmitted given in table 5.5.

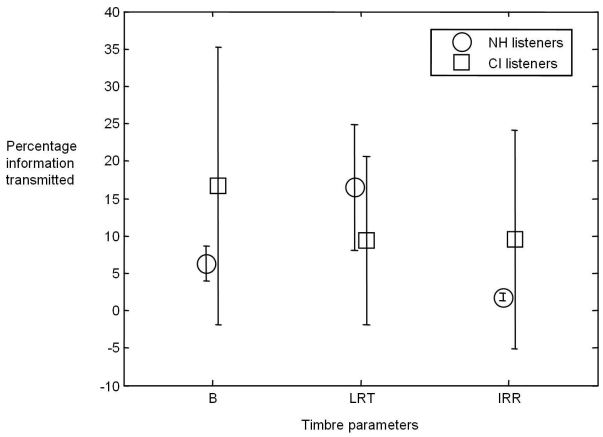

Figure 5.11.
Results obtained from FITA applied to similarity judgements of both NH (circle) and CI (square) subjects for timbre features B, LRT and IRR. NH results are from an average across subjects listening to the processed instrument sounds and CI results are from an average across subjects listening to the unprocessed instrument sounds, with SD values indicated by an errorbar in each case.

Table 5.5.
Averaged FITA mean and SD values for the similarity judgements for both NH and CI listeners.

	NH listeners		CI listeners	
	Mean	SD	Mean	SD
B	6.3132	2.3471	16.6556	18.5775
LRT	16.5159	8.3973	9.3712	11.2350
IRR	1.8244	0.5575	9.5548	14.5965

To determine whether the performance of CI listeners and NH listeners (listening through the acoustic model) differed, a two-factor ANOVA was performed on these experimental results, with the listener type (NH or CI) and the specific timbre feature (B, LRT or IRR) making up the two factors. The effect of the type of timbre feature on the percentage information transmitted by the instrument sounds was found to be non-significant ($F(2,24) = 1.161$, $p > 0.05$), indicating that similar amounts of information were conveyed by each timbre feature. The effect of the type of listener (NH or CI) on the percentage information transmitted was also found to be non-significant ($F(1,24) = 0.785$, $p > 0.05$), implying that NH and CI groups performed similarly in the task. This indicates that the acoustic model was a satisfactory representation of what CI listeners hear, although it was not an exact replica and only included the effect of the CI processor on the sound. There was also a non-significant interaction effect between the listener type and the timbre feature on the percentage information transmitted to the listener ($F(2,24) = 1.755$, $p > 0.05$). This would indicate that NH and CI listeners were not affected differently by different timbre features. However, large SDs in the data may have affected the statistical analysis, where results were found to be non-significant even though there were clear differences in the mean values being compared. For example, differences in the trends of the data points for NH and CI listeners can clearly be seen in figure 5.11, where the three data points for CI listeners form a V-shape, with LRT being a minimum data point, while for NH listeners an opposite trend can be noted, with LRT being a maximum data point.

The percentage information transmitted to NH (mean = 6.31, SD = 2.35) and CI (mean = 16.66, SD = 18.58) listeners for B were somewhat different, with CIs having a higher mean and also a much larger SD. Generally, a much larger SD was observed in the amount of information transmitted to CI listeners, pointing to more uncertainty amongst the CI listeners, or less stable representations of these timbre features in the electrically evoked space-time

action potential patterns. The information transmitted to NH (mean = 16.52, SD = 8.4) and CI (mean = 9.37, SD = 11.24) listeners for LRT showed a different trend from B, with NH listeners having a higher mean than CI listeners, and both groups having large SDs. The information transmitted to NH (mean = 1.82, SD = 0.56) and CI (mean = 9.56, SD = 14.6) listeners by IRR showed a similar trend to that of the information transmitted by B.

5.3.2.2 FITA results of predicted and experimental data

The FITA results as predicted from the timbre perception model, and those measured from the similarity rating experiments, are shown for individual subjects in figure 5.12. Figure 5.12(a) shows the predicted and measured FITA results for each of the NH listeners (NH1 - NH5) in response to the original or unprocessed musical instrument sounds. These results were compiled as a baseline to which the FITA results of the NH listeners listening to the processed sounds and the CI listeners listening to the original sounds could be compared. These results also illustrate the timbre perception model predictions compared to the measured data for NH conditions, the premise on which the model was based.

Figure 5.12(b) shows the predicted and measured FITA results for each of the NH listeners (NH1 - NH5) in response to the musical instrument sounds as processed through the acoustic model. Figure 5.12(c) shows the predicted and measured FITA results for each of the CI listeners (CI1 - CI5) in response to the unprocessed musical instrument sounds.

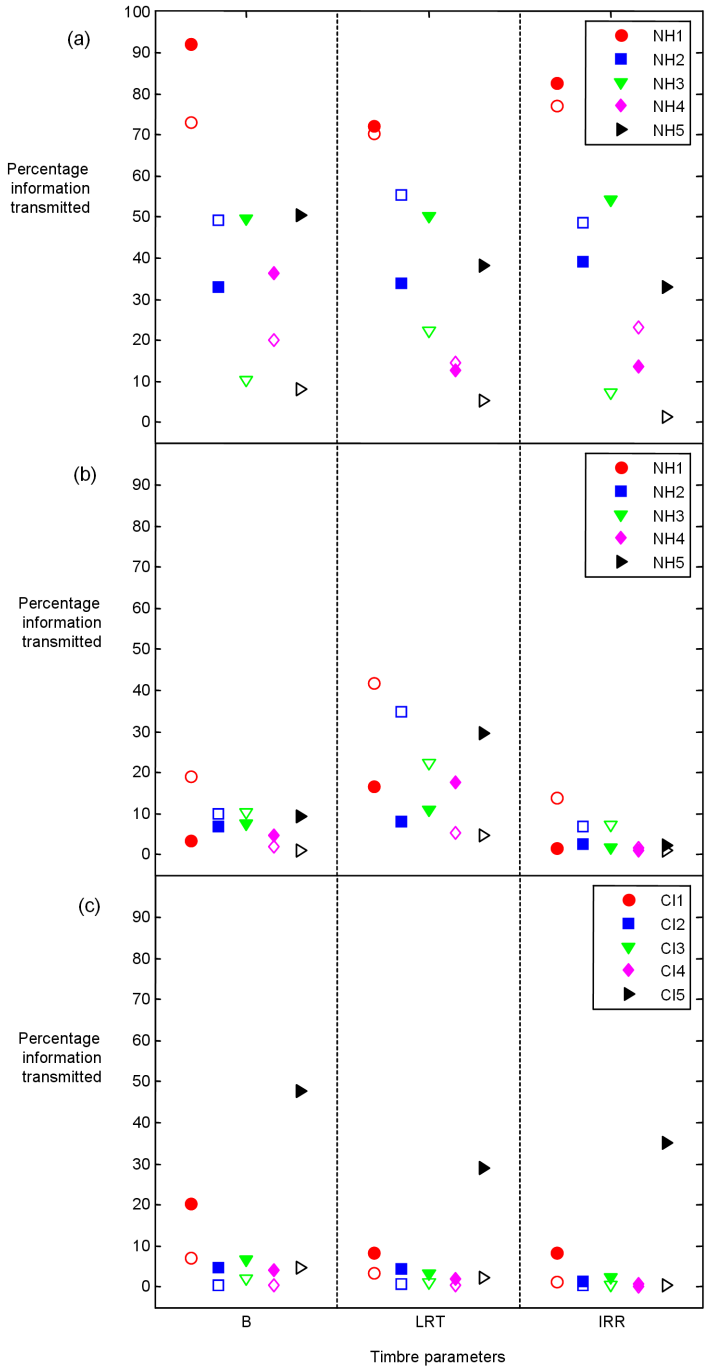

Figure 5.12.
Predicted (unfilled markers) and measured (filled markers) FITA results for (a) each of the five NH subjects in response to the unprocessed instrument sounds, (b) each of the five NH subjects in response to the processed instrument sounds, and (c) each of the 5 CI subjects in response to the unprocessed instrument sounds.

Figure 5.12 illustrates that the timbre perception model fares well in predicting the outcomes of the similarity rating experiments for individual subjects. Although the percentage information transmitted found from the FITA calculations differs substantially across subjects, the model follows the trends of the outcomes of timbre perception experiments for individual subjects. Only in a few specific cases do the model predictions not fare well. For example, in figure 5.12(c), the amount of information transmitted through the timbre features from the measured results for CI5 is very high compared to the other CI listeners, as well as compared to the model predictions for CI5. With the exception of such cases, the model of timbre perception can be seen to generally predict the outcomes of the timbre perception experiments for individual subjects acceptably.

The pooled results of figure 5.12 are shown in figure 5.13, where the averaged FITA results of both the similarity rating experiments and the predicted results obtained from the model of timbre perception for both NH and CI listeners are displayed. Figure 5.13 shows the averaged FITA results for (a) all NH subjects listening to the unprocessed sounds, (b) all NH subjects listening to the sounds processed through the acoustic model, and (c) all CI subjects listening to the unprocessed sounds. Large SDs in the measured data in figure 5.13 can be noted due to the large variations of the FITA results within the subject groups. The mean and SD values are given in table 5.6.

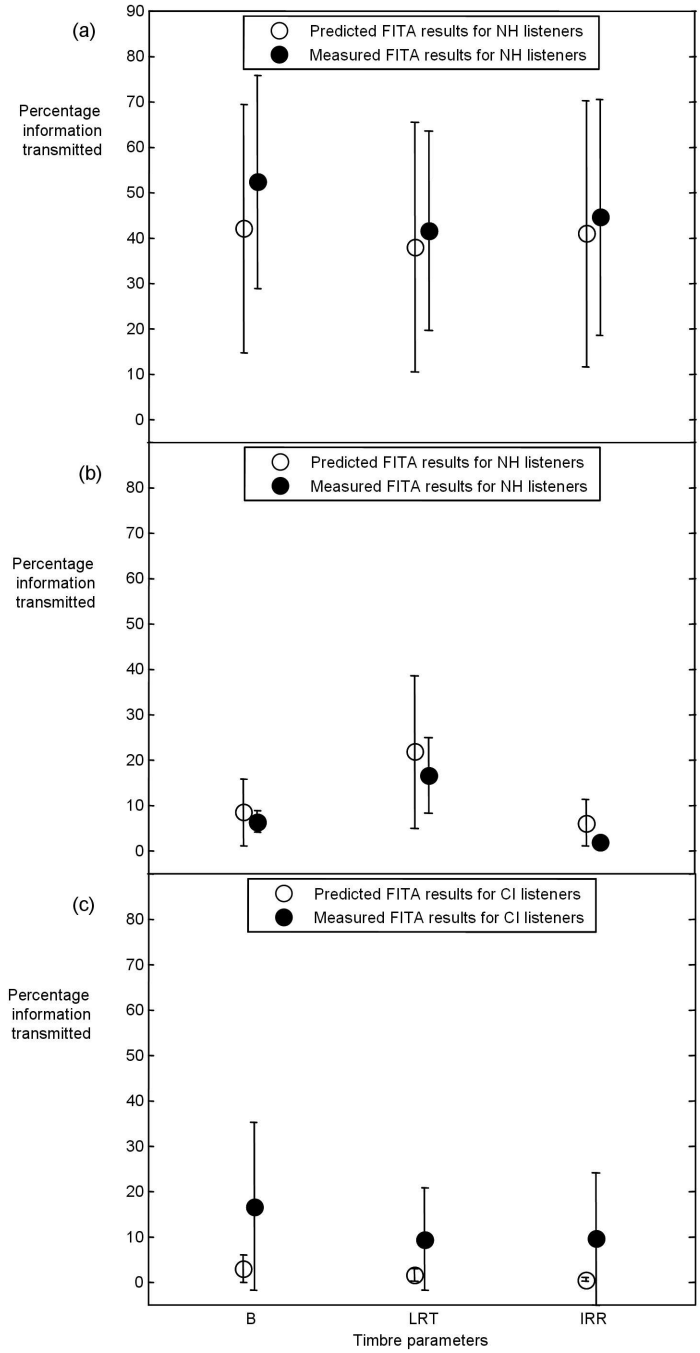

Figure 5.13.
FITA results for both predicted (unfilled circles) and measured (filled circles) data for timbre features B, LRT and IRR for (a) NH listeners subjected to the unprocessed sounds, (b) NH listeners subjected to the processed sounds, and (c) CI listeners subjected to the unprocessed sounds.

Table 5.6.
Averaged measured and predicted FITA mean and SD values for NH and CI listeners.

		Predicted			Measured		
		B	LRT	IRR	B	LRT	IRR
NH listeners	Mean	41.962	37.859	40.850	52.246	41.407	44.489
	SD	27.243	27.482	29.343	23.563	21.932	25.883
NH listeners	Mean	8.4309	21.742	6.083	6.313	16.516	1.824
(processed sounds)	SD	7.391	16.843	5.188	2.347	8.397	0.558
CI listeners	Mean	2.914	1.518	0.481	16.656	9.371	9.555
	SD	2.983	1.263	0.4197	18.578	11.235	14.597

Figure 5.13(a) shows that for each of the timbre features B, LRT and IRR, the percentage information transmitted to NH listeners through the unprocessed instrument sounds is higher for the measured data (filled circles) than for the timbre perception model predictions (unfilled circles). However, the predictions compare well to the measured data, with the largest prediction errors occurring for feature B. As the results of figure 5.13(a) were only calculated from a pilot study of three experimental runs per subject, these are only presented to illustrate the sufficiency of the model predictions in NH conditions. Statistical comparisons will thus only be performed for the results of figure 5.13(b) and (c), as the focus of this study is to investigate timbre perception in the electrically stimulated auditory system.

It can be noted that the predicted mean values (unfilled circles) for each of the timbre features for NH listeners in response to the processed instrument sounds (figure 5.13(b)) are all higher than those predicted for CI listeners (figure 5.13(c)). This would be expected due to the higher JND values obtained by CI listeners than NH listeners, as shown in figures 4.7 and 4.8 for timbre features B, LRT and IRR, on which the predictions are based. However, the NH predictions have substantially larger SDs than the predicted values for CI listeners. The effect of the comparable SD values for the JNDs obtained for NH and CI listeners was more prominent in the predictions calculated for the NH listeners. This is due to the NH predictions being based on the processed B, LRT and IRR values, which are situated in close proximity in the timbre space, as opposed to the original sounds on which the predictions for CI listeners were made. The SDs of the JND values thus produce large SDs in the predictions for NH listeners.

Predicted and experimental values for NH listeners listening to the sounds processed through the acoustic model show that LRT is the feature that conveys the most information, which was expected due to the adequate transmission of temporal information generally found in CIs compared to the limited transmission of spectral information. However, LRT is also the feature with the largest SD, indicating that the LRT feature played a less important role in defining the timbre of a sound for some NH subjects than others.

For CI listeners, the mean predicted percentage transmitted information for each timbre feature is smaller than that found in the experiments. The predicted values also have small SDs, which can again be explained by the SDs of the JND values obtained. These values, as used in the model of timbre perception, have less effect than on the predictions of NH listeners, due to the larger distances between the sounds in the timbre space. The experimental data for each timbre feature exhibit large SDs in comparison, showing varying abilities to perceive timbre amongst the CI subjects. In the case of CI listeners, both the model predictions and experimental data indicated that B was the feature that conveyed the most information on timbre. This contradicted expectations that temporal information, such as the feature LRT, would be transmitted more effectively than spectral information, such as the feature B, as is generally found in CIs. IRR was found to transmit slightly more information than LRT for CI listeners, contrary to the timbre model predictions, in which LRT was found to convey more information than IRR, as would be expected.

A mixed ANOVA was performed on the results of figure 5.13(b) and (c), with the listener type (NH or CI) as the between-subject factor, and the specific timbre feature (B, LRT, IRR) and result type (predicted or real) as the within-subject factors, to analyse individual effects of and interactions between these factors. The type of timbre feature was found to have a significant effect ($F(2,16) = 24.336$, $p < 0.001$) on the amount of information transmitted to the listener. Statistical contrasts revealed that a significantly higher percentage of information was conveyed through the timbre feature B than through IRR ($p = 0.001$), and that a significantly higher percentage information was transmitted through LRT than through IRR ($p < 0.001$), implying that IRR is not generally received by the listener.

The predicted and measured results were not found to be significantly different ($F(1,8) = 0.596$, $p > 0.05$), indicating that the timbre perception model predicted the experimental results sufficiently. Tests of between-subject effects showed that the differences between the results of NH and CI listeners were also not significant ($F(1,8) = 0.75$, $p > 0.05$), again suggesting that the acoustic model was a sufficient representation of sound processing through a CI.

The percentage information transmitted through the different timbre features differed significantly for NH and CI listeners ($F(2,8) = 31.864$, $p < 0.001$), as can be noted from the differences in the trends for NH and CI listeners in figure 5.13(b) and (c). To break down this interaction, within-subject contrasts were performed, comparing different timbre features across NH and CI listeners. These revealed significant interactions when comparing the percentage information transmitted to NH and CI listeners through LRT and IRR ($F(1,8) = 49.061$, $p < 0.001$). The interactions when comparing the percentage information transmitted to NH and CI listeners through B and IRR were not significant ($F(1,8) = 0.647$, $p > 0.05$). Figure 5.13(b) and (c) show that the most information was conveyed by LRT in NH listeners (for predicted and measured results), compared to a substantially lower amount of information conveyed through LRT in CI listeners.

The amount of information transmitted in the different types of results (predicted or measured) did not differ significantly between NH and CI listeners ($F(1,8) = 2.93$, $p > 0.05$). In the predicted results, NH listeners had a substantially higher mean than CI listeners. In the measured results, NH listeners had a lower mean percentage information transmitted than CI listeners, but the difference between NH and CI listeners for measured results was smaller than for predicted values. This suggests that in reality, more information is transmitted to CI listeners than predicted by the timbre perception model. For features B, LRT and IRR for CI listeners, the difference in the percentage information transmitted between predicted and measured values was 13.743 %, 7.853 % and 9.073 %, respectively. For NH listeners, the model predicted that a slightly greater percentage of information would be transmitted than what was actually perceived by the listener. For features B, LRT and IRR for NH listeners, the difference in the percentage information transmitted between predicted and measured values was 2.117 %, 5.226 % and 4.258 % information transmitted, respectively. These results could be due to simplifications made in the implementation of the model of timbre perception as well as in the acoustic model, as will be discussed in section 5.4.

No significant interactions between the different timbre features B, LRT and IRR, and the type of result, predicted or real, were found ($F(1,8) = 0.954$, $p > 0.05$). To break down this interaction, within-subject contrasts were performed, comparing the two levels of result type, predicted or real, across each of the timbre features, B, LRT and IRR. The first within-subject contrast revealed a non-significant interaction ($F(1,8) = 11.622$, $p > 0.05$) when comparing B to IRR when the listener's results were predicted, compared to measured results. This indicates that when comparing B to IRR, there was no difference between the percentage information transmitted between predicted and measured results. There is a decrease in the transmission of information between B and IRR for measured results and a very slight increase in the percentage of information transmitted between B and IRR for predicted results. The means that the measured results for both B and IRR are always higher than the predicted results.

The second within-subject contrast showed a non-significant interaction ($F(1,8) = 0.064$, $p > 0.05$) when comparing LRT to IRR when the results of the listener were predicted, compared to measured results. This shows that there is a decrease in the percentage of information transmitted between LRT and IRR when comparing predicted and measured results. The mean of the predicted results is always lower than that of the measured results for each of the timbre features B, LRT and IRR.

Finally, the interaction effect between timbre features, result type and listener was not significant ($F(2,16) = 0.102$, $p > 0.05$) for interactions between B and IRR and between LRT and IRR. This indicates that the interaction between timbre feature and type of result was not different for NH and CI listeners. This in turn suggests that, overall, for predictions and measured results, the acoustic model performed well. Again, contrasts were used to break down the interaction: these contrasts compared the information transmitted to NH and CI listeners at both predicted and measured result levels across each of the timbre features. The first contrast revealed a non-significant difference ($F(1,8) = 1.602$, $p = 0.241$) between NH and CI listener values when comparing B to IRR for predicted compared to measured results, and tells us that for both NH and CI listeners, there is a decrease in the percentage of information transmitted between B and IRR for both predicted and measured results. In the case of NH listeners, the predicted mean results are always higher than the measured results, whereas the opposite applies to CI listeners.

The second contrast investigated differences between NH and CI listeners when comparing LRT to IRR for predicted compared to measured results. This contrast was found to also have a non-significant interaction effect ($F(1,8) = 0.001$, $p = 0.977$). This shows that there is a decrease in the percentage information transmitted between LRT and IRR when comparing predicted and measured results, for both NH and CI listeners. The mean of the predicted results is substantially lower than that of the measured results for each of the timbre features B, LRT and IRR for CI listeners, while there are smaller differences between predicted and measured results for NH listeners for all of the timbre features, but with the predicted means being slightly higher than the measured means in each case.

Although there are differences in the trends noted for NH and CI listeners, the predicted and measured results follow the same trend for each listener group, as shown in the results of figure 5.13. This shows that the model of timbre perception provides an adequate representation of timbre perception, defined by B, LRT and IRR, and shows that the amount of information transmitted by these features through a CI is indeed low (see figures 5.13(b) and (c), compared to the NH conditions of (a)). This suggests that improvements in conveying B, LRT and IRR to the listener may be the key to improving timbre perception in the electrically stimulated auditory system.

5.4 DISCUSSION

5.4.1 FITA analysis of similarity ratings

The FITA results of figure 5.11 illustrate the differences in the similarity rating results for NH listeners (listening to the acoustic model) and CI listeners. Although the trends for the timbre features differ for the two listener groups, the amount of information transmitted by the timbre features is in a comparable low-end range for both NH and CI listeners. This suggests that the implemented acoustic model fares sufficiently in comparison to the results obtained from CI listeners, but that the acoustic model is not an exact replica of what CI listeners hear. This was expected, as the ability of CI listeners to perceive sounds varies, and therefore a good general acoustic model should be able to predict the average outcome across a large group of listeners, as opposed to providing accurate predictions for individual listeners. The outcome of the acoustic model as implemented for the purpose of this study is

therefore sufficient, as the primary concern was not with designing a good acoustic model, but rather predicting the timbre data found in NH and CI listeners sufficiently, in order to determine whether timbre perception could be modelled using the three important timbre features as a foundation.

The ranking of features through which the highest percentage information was transmitted differs for the NH and CI groups. For the NH group, the most information was transmitted through the feature LRT, followed by B, and then IRR. This is what we would expect from NH listeners exposed to acoustically modelled sounds, as the temporal information is least affected by processing of the sound, as discussed in section 4.4.1.1. However, for the CI group, the least information was transmitted through LRT, with B transmitting the greatest amount of information, followed by the IRR feature. This would indicate that on average, the CI listeners have a better perception of the spectral features of timbre than is assumed by the acoustic model, and a worse perception of the temporal features of timbre than is assumed by the acoustic model. This finding can in part be substantiated by a study conducted by Stainsby *et al.* (2002), in which steady-state envelopes of musical instrument sounds were investigated. The study showed that some CI users may have frequency selectivity that is comparable to that of NH listeners, and also concluded that a large amount of spectral information seems to be available to some CI listeners. This finding could explain why the feature B is transmitted most effectively to CI listeners in the similarity rating experiments of this study. However, due to the large SDs in the CI group of this study, as well as those found by Stainsby *et al.* (2002), it is difficult to make general conclusions, because subject-specific factors affect individual results differently.

NH listeners generally received the most information through the LRT feature, due to the structure of the acoustic model, but it is also the feature in the NH group with the largest SD, indicating substantial variations in the results of the NH subjects for the perception of this feature.

The similar trends between features B and IRR for the information transmitted to NH and CI listeners can be explained by both of these features involving the spectral composition of the instrument sounds. The subject-specific spectral resolution of CI listeners is illustrated by the large SDs in the amount of information transmitted to CI listeners for features B and IRR. The mean values for CI listeners for features B and IRR are higher than for NH listeners, but the NH group has small SDs for both features. This is indicative of the strict limitations imposed on the spectral components of the sound through the acoustic model,

which cause the percentage information transmitted to be less than in the CI listeners. These limitations also cause smaller SDs in the NH group for B and IRR, as the spectral limitations substantially affect the timbres of the sounds, which are clearly conveyed to the listener and therefore perceived similarly in all listeners.

5.4.2 Model of timbre perception

To assess the success of the developed model of timbre perception, a comparison of the predicted and measured confusion matrices obtained for both NH and CI listeners will first be discussed. The predicted NH and CI confusion matrices of section 5.3.2.2, as shown in figures 5.9 and 5.10, show similarities in the instrument confusions. Specifically, the piano sound is least confused with any of the other sounds for both NH and CI listeners subjected to the processed and unprocessed sounds, respectively. In both predicted confusion matrices, the saxophone is most likely to be confused with the piano. Higher confusions can be noted for both NH and CI listeners between instruments in the family of string instruments. These trends also correspond to the confusion matrices obtained from the similarity rating experiments.

Similarities between the predicted confusion matrices also include a high probability of confusing the flute and cello, amongst other higher confusions between sound pairs. Overall, the predictions for NH and CI listeners appear to have higher confusions that are more scattered throughout instrument pairs than in the case of the experimental results of figures 4.9 and 4.10. This is possibly due to the fact that the JND values used in creating the ellipsoids around each instrument sound in the timbre space were found from the averaged discrimination of synthesised tones. This could have contributed to making the predicted confusions higher and more scattered than was found for the measured abilities of the listener.

Comparisons between the predicted and measured FITA results for both NH and CI listeners, as given in figures 5.12 and 5.13, illustrate the outcomes of the developed model of timbre perception. The individual listener results of figure 5.12 show that although the percentage of information transmitted to each listener varies substantially across subjects, the trends of the percentage of information transmitted through each timbre feature are similar for the predicted and the measured data for each listener, with only a few exceptions. For example, in figure 5.12 (c), subject CI5 is seen to perform comparably to NH listeners (figure 5.12 (a)) in the similarity ratings, but the predicted results underestimate the performance of this subject

substantially. However, in addition to the fact that CI5 had been exposed to musical training, this subject also enjoyed listening to music and showed exceptional speech perception abilities upon receiving her implant, suggesting extraordinary overall music perception abilities which could not be accurately predicted by the model.

However, the similarities generally found in the predicted and measured percentage of information transmitted for each subject suggest that the model of timbre perception sufficiently predicts the measured data for each listener, illustrating the subject-specific nature of the model. This outcome also suggests that the method of modelling ellipsoids around each of the instrument sounds in the three-dimensional space to obtain the predicted data was appropriate, as individual listener JNDs were used to predict confusions between instrument sounds. If Euclidean distances alone had been used for predicting the confusions between instrument sounds, the predictions within each listener group would have been identical and the model would not have predicted the outcomes of the similarity ratings for each subject sufficiently.

The FITA analysis of figure 5.13 revealed close predicted values compared to those measured experimentally for NH listeners (figure 5.13 (a) and (b)) for each of the timbre features. However, fairly large differences in predicted and measured results for each of the timbre features were found for CI listeners. Large SDs in the FITA results are apparent from the individual subject results in figure 5.12, where the variability in the results among subjects was large.

From the results of figure 5.13 (b), the NH listener predictions, it can be seen that the predicted percentage of information transmitted by each of the timbre features was always slightly higher than the percentage of information transmitted as calculated from the measured results. This could be due to simplifications made in the implementation of the model of timbre perception, as the JND values used in the model were obtained from simplified instrument sounds, constructed only from the timbre features B, LRT and IRR. The omission of factors such as noise in the synthesised sounds from which the JNDs were calculated may have caused the model to overestimate the percentage of information transmitted through each of the timbre features, as real instruments sounds were used in the similarity ratings.

As a result of not using real sounds for the discrimination tasks, differences in JND values for timbre features and the percentage of information transmitted through these features for some of the listeners can be noted. For example, subjects CI2 and CI3 have fairly low B

JNDs (figure 4.7), but a poor percentage of information transmitted for B (figure 5.12). Similarly, subject CI5 had an average LRT JND, but a high percentage of information transmitted through LRT. These discrepancies indicate that using real sounds for the discrimination tasks could change the outcomes of the discrimination tasks substantially, and provide better similarities between the predicted and actual outcomes for the model of timbre perception.

The processing of the real instrument sounds through the acoustic model possibly further increased the complexity of the sounds, and thus the timbre similarity rating tasks, in comparison to the discrimination tasks, for the listener. In the model of timbre perception, the gain factor, as discussed in section 5.2.1 for B, LRT and IRR, could perhaps be adjusted to below unity to compensate for the difficulties introduced in perceiving the sounds as a result of processing through the acoustic model and thereby make the timbre perception model predictions more accurate.

For CI listeners (figure 5.13 (c)), the timbre perception model predictions over-estimate the difficulty of the similarity rating task. A possible explanation for this finding may be that CI listeners make use of other information in addition to the NH timbre cues B, LRT and IRR to perceive the timbre of a sound. This could be due to CI listeners having grown accustomed to sounds heard every day through a CI, along with which additional available information may be utilised to perceive auditory stimuli. By using synthesised sounds as opposed to real instrument sounds in the similarity ratings, the possible additional cues would be absent, and this would perhaps result in a better correlation between the measured and predicted results.

In addition, not having included signal pre-emphasis in the acoustic model may explain the resulting low predictions of the model for CI performance. Although the model predicted NH performance fairly well, the model under-estimated the abilities of CI users quite substantially. As a result of excluding signal pre-emphasis in the acoustic model, the stimuli used as a basis for the model as well as in the CI simulation, would have been quite low-pass in nature. However, in the case of real CI listeners, where signal pre-emphasis is included, more higher frequency channels may have been stimulated, perhaps heightening the perceptual capabilities of CI listeners in comparison to the predictions of the model. By including signal pre-emphasis, the percentage information transmitted to NH listeners listening to the processed sounds could increase for both the predicted and measured data. For the results of the CI listeners, including signal pre-emphasis would potentially result in higher predicted percentages of information being transmitted to the listener, thus potentially moving the predicted and measured data closer in proximity, thereby improving the accuracy of the model.

Differences in the predicted and measured results obtained for CI listeners may also have been strongly dependent on the acoustic model implementation being different to the processing as performed through a real CI. The biophysical characteristics of the electrode-neural interface were not included in the developed acoustic model, but in the case of CI listeners, the biophysical aspects were present. The omission of the biophysical characteristics from the acoustic model may also explain why there were differences between the NH and CI results - for predicted and measured FITA results. This is as a result of the different processed sounds that the NH and CI listeners would be exposed to, with NH listeners having no biophysical effects or signal pre-emphasis, while in the case of CI listeners, these factors are utilised in the processing of the musical instrument sounds.

However, the omission of the biophysical characteristics of the electrode-neural interface allowed the effect of only the processor on timbre perception to be investigated. If the processor had not had an influence on timbre perception, this would not have been apparent by implementing both the processing and biophysical aspects of the acoustic model as a first acoustic model implementation for this study. As can be seen from the differences in the frequency domain representations of figures 3.33 and 3.34, the effect of the processor as implemented in the acoustic model had a drastic effect on the frequency spectra of the sounds. By comparing these figures, the frequency peaks of the processed sounds in figure 3.33 are all low pass in nature. This is due to the implementation of the ACE algorithm; because the energy is concentrated in the lower frequency components for the musical instrument sounds, the lower frequencies are selected.

By including the biophysical characteristics in the acoustic model implementation in future revisions of this work, differences between the NH and CI FITA results of figure 5.13 may be decreased. Differences between the predicted and measured results of the CI FITA results could be decreased by including the biophysical characteristics in the acoustic model implementation, but the exact effect can not be known without conducting further experiments to establish this.

Overall, it is difficult to make substantiated comparisons between the trends of NH and CIs, probably due primarily to the limitations in the predictive abilitiy of the acoustic model, as seen from the measured data in figure 5.11. From these results, it can be expected that differences between NH and CI model predictions would occur. An acoustic model with the biophysical characteristics of the electrode-neural interface included may yield comparable trends in NH and CI data, implying that NH listeners could be used to predict the outcomes

of timbre perception experiments for both NH and electrically stimulated hearing conditions. However, this is not a primary concern, as the objective here was not to develop an acoustic model that could predict timbre perception. Rather, the NH data serves as a baseline for comparison, while the main objective was to develop a model that would be able to predict CI timbre perception data. The present work provides a foundation for this purpose, showing that the modelling ideas are correct (as evidenced by figure 5.13 (a) and (b)), but still have shortcomings (as evidenced by figure 5.13 (c)).

A likely explanation as to the sufficient predictive abilities of the model of timbre perception for NH listeners compared to CI listeners can be provided by the choice of the important timbre features; namely, B, LRT and IRR. These features have been reported in literature as important features for NH listeners, but no features have been explicitly defined for CI listeners. Since we do not explicitly know what the dimensions of timbre are that facilitate timbre perception in CI listeners, we based the model on NH features, and could expect that this model would not represent timbre perception in CI listeners as well as in NH listeners. As suggested previously, CI listeners may use other auditory cues to perceive timbre in addition to the three important NH cues. The hypothesis that CI listeners make use of other auditory cues to perceive timbre would need to be tested to draw further conclusions regarding the differences in the performance of the model of timbre perception for NH and CI listeners.

Additionally, a factor that may have contributed to the poorer predictive abilities of the model of timbre perception for CI listeners was the large variations in the abilities of the CI listeners to perceive timbre, specifically for CI5, as discussed previously (figure 5.12 (c)). Only a small subject group was used in this study, whereas average model predictions and measurements of perception calculated over a larger subject group may yield more consistent results.

From figure 5.13, it appears that the model of timbre perception is sensitive to the position of the sounds in the three-dimensional timbre space. This can be illustrated by comparing the NH predicted results of figure 5.13 (a) and (b), for unprocessed and processed instrument sounds, respectively. Processing of the sounds through the acoustic model causes the sounds to be shifted substantially in the timbre space, while the JNDs used in obtaining the predictions remain the same for NH listeners in both figure 5.13 (a) and (b). However, the predictions for these two cases differ substantially, also showing the sensitivity of the model of timbre perception to the choice of the acoustic model implemented, as this will affect the positioning of the sounds in the timbre space.

The model sensitivity may thus also be affected by factors such as the exclusion of signal pre-emphasis from the developed acoustic model, which could have affected the grouping of sounds in the timbre space and thus could have had a substantial effect on the predictions of the model of timbre perception, due to the model sensitivity to the position of the sounds in the timbre space. Additionally, the exclusion of the biophysical characteristics from the acoustic model could also have greatly affected the sensitivity of the model of timbre perception, as this would have affected the positions of the sounds in the timbre space.

The SDs indicated for the predicted results of figure 5.13 are due to the differences in the JNDs obtained for each subject. This illustrates the sensitivity of the model of timbre perception to the discrimination task results and therefore to individual subjects, suggesting that the model is sensitive to subject-specificity.

A general trend that can be observed from the average results of figure 5.13 is a relative increase in the SD as the percentage of information transmitted increases. This is indicative of the sensitivity of the model of timbre perception to both the position of the instrument sounds in the three-dimensional timbre space and to subject-specificity, and also suggests that the JND values and the position of the sounds have a relative impact on the model predictions.

A summary of the differences in the results of figure 5.13 (b) and (c) for predicted and measured results for NH and CI follows. For NH listeners, LRT was predicted to be the feature that would be conveyed most readily to the listener, followed by a prediction of a much lower amount of information transmitted by B, and then by IRR. The experimental results showed the same trend. In addition to the predicted outcomes following the same pattern as the measured results, the predicted and measured results showed on average less than a 3 % difference in percentage information transmitted.

However, for CI listeners, predictions and experimental results both indicated that B would be the feature conveyed most readily to the listener. In the predicted results, B was followed by LRT, and then by IRR being conveyed least effectively, while in experimental results IRR conveyed slightly more information than LRT. In addition to these differences in trends, the differences in predicted and measured results were greater than for NH listeners, with an average of over 10 % difference in percentage information transmitted.

Overall, the FITA results of figures 5.12 and 5.13 illustrate the general sufficiency of the timbre perception model. The model appears to predict the outcomes of the timbre similarity rating experiments for NH and CI listeners acceptably, given that the trends of the predicted results follow those of the measured results. In addition, the trends of figure 5.13 (a) and (c), for NH and CI listeners respectively listening to unprocessed sounds, are similar, but the NH listeners clearly have a higher percentage of information available to them through the three timbre features. These similarities in the trends appear to confirm the validity of the choice of B, LRT and IRR as the primary contributing features to timbre perception.

The results suggest that, if CI processors could be optimised for the transmission of these three important timbre perception features, timbre perception through a CI should improve. Alternatively, these three features could be used as a relative measure when comparing new speech processors that may be designed to improve timbre perception. This approach could provide a favourable alternative to memory-based tasks, such as instrument identification, which are commonly used in timbre perception research.

5.5 CHAPTER SUMMARY

This chapter described the modelling component of this study in detail. The timbre features B, LRT and IRR, extracted for both the unprocessed instrument sounds and those processed through the acoustic model, as well as the JNDs for these values obtained from the discrimination task results (section 4.3.2), were provided and formed a foundation for the development of the model of timbre perception. The predicted results of the model of timbre perception were reported in the form of confusion matrices. FITA analyses were performed on these results as well as the confusion matrix results of the similarity rating experiment (section 4.3.2). This analysis indicated the percentage information transmitted through each of the important timbre features for predicted and measured results for both NH and CI listeners. Statistical analyses of the results were also provided. A detailed discussion of the results of the model of timbre perception were presented with comparisons made to literature where possible. This chapter discusses the overall outcomes and implications of this study, which provides an entry point into achieving a quantitative understanding of the timbre perception abilities of CI listeners.

The model of timbre perception developed based on the timbre features B, LRT and IRR has been shown to adequately predict the outcomes of timbre perception experiments for both NH and CI listeners. This provides a valuable tool in developing CI processors to facilitate timbre perception, and thus in furthering timbre perception research, with the ultimate aim of improving timbre and music perception for CI listeners. A general discussion and conclusion is provided in chapter 6, to summarise the accomplishments of the study and to provide a critical analysis with directives for future work, using this study as a foundation.

CHAPTER 6

GENERAL DISCUSSION AND CONCLUSION

A brief general discussion of the measurement and modelling components of this study, presented in chapters 4 and 5, is given below. Detailed discussions of the work presented in these chapters are given in sections 4.4 and 5.4. A brief summary is given here to illustrate that the research questions posed in chapter 1 have been addressed.

- The important timbre perception features were defined as B, LRT and IRR (section 3.4), and were successfully extracted from both the original sounds and sounds processed through the acoustic model (section 5.3.1).
- Although the acoustic model did not predict the outcome of CI timbre perception experiments accurately, as shown in section 5.3.2.1, the model was acceptable for the purpose of this study, in which the primary focus was not on the acoustic model.
- Quantitative results were obtained regarding timbre perception in NH and CI listeners, presenting the abilities of the perception of the important timbre features B, LRT and IRR in measurable terms for both listener groups (section 4.3.1).
- The model of timbre perception developed sufficiently predicts the results of NH and CI listener timbre perception experiments, with the trends of the predictions following those of the experimental results in both groups, as can be seen from figures 5.13 in section 5.3.2.2.

The main findings of this study are listed below.

- Measurable results for three important timbre features, namely B, LRT and IRR, were found for NH and CI listeners (section 4.3.1), and compared well to NH and CI literature regarding overall timbre perception abilities. NH listeners showed substantially better discrimination abilities than CI listeners for each of the timbre features, as can be seen from figure 4.7 in section 4.3.1.

- From figure 4.8, the CI listeners were seen to be most sensitive to the temporal feature LRT. Discrimination of the spectral centroid showed large variations among subjects of this group, and a poor discrimination ability in general was observed when compared to NH listeners.

- The developed acoustic model did not provide an accurate representation of timbre through the electrically stimulated auditory system. The results of figure 5.11 in section 5.3.2.1 show that the suppression of the spectral features was far greater through the acoustic model than what was actually perceived by the CI subjects, while the temporal feature limitations imposed were not great enough. The large SDs in the amount of information transmitted through each timbre feature to CI listeners illustrates the subject-specific nature of CI processors, as substantial differences were found from subject to subject (see figure 5.12).

- The timbre perception model predicts the transmission of timbre features to NH and CI listeners satisfactorily, as can be seen from the results of figure 5.13. For NH listeners, the model of timbre perception provides accurate predictions, with an approximate average difference of less than 4 % of the information transmitted between predicted and measured results across all of the timbre features. The predicted amount of information transmitted through each timbre feature was always slightly higher than the measured amount transmitted to the listener as found by the similarity judgements of timbres. This indicated that the model predictions overestimated the abilities of NH listeners to perceive each of the timbre features through the acoustic model. In CI listeners, the model of timbre perception performs less accurately, with a difference of greater than 10 % between the percentage information transmitted between predicted and measured results incurred over all the timbre features. The model predictions underestimated the abilities of CI listeners to perceive each of the timbre features.

- The predictions of the model of timbre perception indicate that the transmission of the important timbre features B, LRT and IRR through the electrically stimulated auditory system is poor, as can be seen from figure 5.13 when comparing (a) the NH listener results to (c) the CI listener results for unprocessed sounds. This implies that timbre perception abilities of CI listeners will be improved by the development of processing strategies to facilitate the transmission of these features.

The first step of this study was to perform a general literature study on CIs, existing acoustic models and timbre perception for both NH and CI listeners. The literature study was then focussed to investigate timbre features deemed important for timbre perception that could be used as a basis on which to develop a model of timbre perception. The timbre features B, LRT and IRR were established as important timbre features through existing literature to complete the first two objectives of the study, as described in section 1.2. Although these three features were prominently found to be linked to timbre perception in literature (e.g. Krimphoff *et al.* (1994), Caclin *et al.* (2005) and McAdams *et al.* (1995)), other features were however also found to be important. A shortcoming of this study is that not all the features important for timbre perception were included in the study. For example, a recent study by Hall and Beauchamp (2009) found that both spectral incoherence and spectral irregularity were relevant features for the perception of musical instrument tones in NH listeners. This would suggest that including additional timbre features in the model of timbre perception, especially where there are some discrepancies in the literature as to which features are most important, could improve the accuracy of the model predictions.

In addition, this study assumed that the features deemed as important for timbre perception in CI listeners were the same as those for NH listeners, as a result of a lack of literature to suggest otherwise. As could be expected, the predictions of the model showed larger errors for CI listeners than for NH listeners. This suggests that other timbre features important for timbre perception in CI listeners may exist which need to be included in the model of timbre perception for CI listeners. Although the approach of assuming the same features applied to CI listeners as to NH listeners was fair given that studies such as this had not been formulated before, other timbre features could have been considered with the potential of improving the model of timbre perception for CI listeners. An investigation into the acoustic cues used by CI listeners to perceive timbre could be conducted to gain insight into the important timbre features in the electrically stimulated auditory system, a component that was lacking in this study.

Extraction of the important timbre cues for CI listeners through extensive psychoacoustic experiments may be possible by using a similar approach to timbre perception studies for NH listeners as in previous literature, where MDS techniques were commonly utilised (see section 2.3.1). However, investigating timbre features for the electrically stimulated auditory system would present difficulties in that simple instrument recognition tasks could not be applied reliably, with many CI users not having the musical memory required to complete such tasks. Similarity ratings, such as those performed in this study, but with more extensive and complete sets of instruments, could be a possible entry point into the extraction of important timbre features for CI listeners. By attempting to extract the acoustic cues for CI listeners from MDS results extracted from such similarity ratings, it may be possible to develop a model of timbre perception unique to CI listeners using specific CI timbre features as opposed to the NH features B, LRT and IRR implemented in this model.

The literature study allowed for the development of an acoustic model based on the ACE strategy to complete the third objective given in section 1.2. Only the processing side of the acoustic model was implemented, without taking into account the effects of the biophysical characteristics of the electrode-neural interface, to isolate the effects of a CI processor on musical instrument sounds. In addition, signal pre-emphasis was not included in the processing steps of the acoustic model as this study focussed on music, and pre-emphasis is usually included for speech intelligibility. This basic implementation of the acoustic model served as an entry-point into understanding how timbre perception is affected by processing through a CI, and as a basis on which to perform timbre perception experiments with NH listeners. However, experiments involving similarity ratings of timbres, with NH listeners listening to acoustic simulations and CI listeners listening to unprocessed musical timbres, revealed fairly different results for the different listener groups. This indicates that the acoustic model was not an accurate representation of timbre as presented to the electrically stimulated auditory system, but was sufficient for the purposes of this study, as the acoustic model was not the primary focus.

Improvements to the accuracy of the implementation of the acoustic model should be investigated for future revisions of this study. Including signal pre-emphasis could be considered to create a new set of stimuli from which the effects of the CI processor on timbres could be more accurately assessed. In addition, the effect of biophysical characteristics of the electrode-neural interface should be included in the acoustic model to provide a more realistic representation of timbres as processed through a CI, as well as providing an improved foundation on which to model timbre perception. Improvements to acoustic models apply not only to this study, but to all CI research. Acoustic models provide generic representations of sound through a CI but fail to accurately represent what is heard by a variety of implantees. This in turn results in acoustic models not always being able to predict the outcomes of CI experiments, a shortcoming which needs to be addressed before acoustic models can be used to their full potential in CI research. With a more accurate representation of sound as it is processed through the electrically stimulated auditory system, more accurate predictions of the performance of CI listeners could be made and a closer correspondence to the predicted and measured results for NH and CI listeners for the timbre perception model implemented in this study would be possible.

The experimental component of this study consisted of timbre discrimination tasks as well as similarity ratings, using synthesised and real musical instrument sounds, respectively. Although peak normalisation of the sounds used in the experiments was performed and assisted in balancing the perceived loudness levels of the sounds, a subjective loudness balancing procedure was not implemented, as this would require new complex procedures to be developed for such a task. However, the peak normalisation of the sounds used in this study was adequate, as there were no large perceptual loudness differences in the sounds. In future revisions of this work, an option could be to normalise the RMS of the sounds while ensuring that no peak clipping occurs. This would ensure that each of the sounds would have the same total energy, and should eliminate any effect that perceptual loudness difference might have on the similarity ratings. By balancing the energies of the sounds, improved stimuli for both experiments and modelling could result, possibly providing more accurate results for the similarity rating experiments, as well as improving the accuracy of the timbre perception model predictions.

Timbre feature discrimination abilities for both NH and CI listeners were measured by means of psychoacoustic experiments with synthesised sounds. This provided quantitative results of the perceptual abilities of both NH and CI listeners, and differences therein, for the three important timbre features. Synthesised sounds were used for ease of generating and altering

sounds, as used in previous studies regarding timbre perception (e.g. Caclin *et al.* (2005)). Discriminations of the synthesised tones were used to find JND values (as in section 4.2.1) to be used as inputs to the model of timbre perception (based on the method of van Zyl (2008)) that would be common to all sounds for each listener. Although this served as a satisfactory entry-point to gain such experimental data, alternative methods could be investigated to provide more accurate JND values. Instead of using a generalised synthesised tone in the discrimination tasks in which JND values were recorded, a possibility is to use real musical instrument sounds that are varied by adjusting one of the timbre features each time. A study by Horner, Beauchamp and So (2009) followed such an approach for timbre perception studies in NH listeners and found that timbre discrimination abilities differed for changes in different musical instruments. However, a shortcoming of this study was to assume the same JNDs for all instrument sounds for each listener.

A more accurate representation of the JNDs for B, LRT and IRR around individual musical instrument sounds would be possible if real instrument sounds were used, but this would require a computationally intensive and accurate timbre resynthesis technique such as that developed by Jensen (1999b) to be implemented for a variety of musical instrument sounds. The sounds would have to be resynthesised for each incremental change in timbre features, which calls for a complex process requiring highly accurate timbre resynthesis techniques. The number of experimental sessions would also increase substantially, as individual experiments for discrimination abilities of each timbre feature for each musical instrument sound would have to be investigated. However, a database of such measurable results may provide invaluable insight into timbre perception abilities of NH listeners compared to CI listeners.

Shortcomings in this work as a result of not using real sounds for the discrimination tasks can be noted from differences in JND values for timbre features and the percentage of information transmitted through these features for some of the listeners, as discussed in section 5.4.2. Using real sounds for the discrimination tasks in future revisions of the work performed in this study could change the outcomes of the discrimination tasks substantially, and provide better similarities between the predicted and measured outcomes for the model of timbre perception. If the same sounds are used for both the discrimination tasks and similarity ratings, a correlation between timbre feature JNDs and the percentage information transmitted through these features could prove to be useful in further understanding timbre perception, and thus in validating the developed model of timbre perception.

The development of a model of timbre perception was carried out using the timbre feature JND results obtained from the discrimination tasks as a basis to complete the main objective of the study. An approach similar to a model of speech perception previously implemented (van Zyl, 2008), and based on signal detection theory, was followed to provide predictions for timbre perceptual studies which were compared to measured results in the form of similarity ratings. The dimensions of the ellipsoids used in the model of timbre perception were implemented to be symmetrically positioned around the sounds in the three-dimensional space, as described in section 5.2.1. This modelling assumption is one that may not be well substantiated, although it was used in the model by Svirsky (2000) previously. Non-symmetrical ellipsoids could be an option for exploration to obtain more accurate predictions.

The timbre perception model predictions were compared to measured results obtained through similarity ratings of ten musical instrument sounds. The similarity ratings were used as a measure of confusions between musical instrument sounds. Although direct instrument identification tasks would have produced confusion matrices directly, it was decided to use similarity ratings instead, primarily due to the difficulties associated with an identification task for CI listeners. Most of the CI listeners were limited in musical memory, and would not have been able to identify instruments by name. Even the similarity rating tasks were extremely challenging to most of the CI listeners. However, in future work an attempt should be made to make comparisons between confusion matrices obtained from similarity ratings and identification tasks, as the relative weights for timbre perception might assume different proportions in an instrument identification task than in similarity ratings.

A possibility for future revisions of this work could also be to perform similarity ratings without including like pairs of sounds, for example the piano-piano sound pair. This would potentially give rise to different similarity rating results, with the listener making use of a different perceptual weighting scale. In addition, the effect of similarity ratings between mismatched sound pairs could also be more prominent by excluding like pairs of sounds in the similarity rating tasks, possibly resulting in different confusion matrices and thus FITA results.

The instrument sounds used in this study were selected as an adequate sample group containing instruments from the four primary musical instrument families, as discussed in section 3.2. To expand on this study and possibly improve the model of timbre perception, an option would be to encompass a wide range of musical instrument timbres, as found in everyday music listening conditions. This will involve investigation of larger sound databases in addi-

tion to the 10 instrument sounds investigated in this study, as well as variation of the notes of the musical instrument timbres presented. As discussed by Hall and Beauchamp (2009), the identification of some timbres can change depending on pitch, with the possibility of characteristic pitches emerging for instruments. A shortcoming of the research in this dissertation was that only C4 (262 Hz) notes were used, but other common notes used in western music and timbre perception studies include F#3 (185 Hz), E4 (330 Hz) and G4 (391 Hz) (Nimmons *et al.*, 2008), and these may provide insight into attributes of timbre perception that are not visible when using only a single pitch.

Additionally, potential training effects may have influenced the results of both the discrimination and the similarity rating tasks, as a small closed set of stimuli were used. Training effects could thus also have affected the differences between the predicted and measured CI performance. In future revisions of this work, a more complete set of stimuli could be developed, encompassing different pitches and timbres, with the aim of presenting stimuli in a complex environment more true to real listening conditions.

Through analysis of the confusion matrices obtained from both the model predictions and similarity rating experiments conducted, it was possible to establish which perceptual timbre features are available and to what extent they are conveyed through the processor of a CI. The results of the timbre perception model showed that the model sufficiently predicted the results of the timbre similarity rating experiments for NH and CI listeners, with the trends of the predicted values following those of the measured results. For NH listeners, the model predictions were fairly accurate, with only small discrepancies between the predicted and measured results obtained. However, for CI listeners, the differences between predicted and measured values showed larger discrepancies, and this was probably due to the timbre features on which the model was based being NH timbre features. In general, the model of timbre perception was found to predict the outcomes of timbre perception experiments satisfactorily for both NH and CI listeners, thereby providing a tool to assist in the development of CI processors to facilitate timbre perception, and thus accelerate CI research.

Because the research addressed in this study is a recent topic, measurable results regarding timbre perception, particularly in CI listeners, are not readily available. More extensive investigation into both discrimination abilities of timbre features and similarity judgements of musical instrument timbres is a suggestion for future work necessary to gain a better understanding of timbre perception abilities in both NH and CI listeners in general.

The next research step will be to improve CI processors to assist with the transmission of important timbre features to the listener. Improving the transmission of the important timbre features B, LRT and IRR by means such as those suggested in section 4.4.1 could enhance timbre perception in CIs. Investigation into additional cues used by CI listeners, and the effective incorporation of these to be optimally transmitted through a CI, could also be beneficial to advancing timbre perception in the electrically stimulated auditory system, and indicates the vast research opportunities that can stem from this study.

The initial model of timbre perception presented in this study provides a tool to assist in CI research regarding timbre perception by providing a quantitative understanding of timbre perception in the electrically stimulated auditory system. In turn, this provides a platform on which to develop CI processors to facilitate timbre perception, with the ultimate goal of improving music perception through a CI processor.

REFERENCES

Adank, P., Smits, R. and Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research, *Journal of the Acoustical Society of America* **116**(5): 3099–3107.

American Standards Association (1960). *Acoustical terminology*. S1.1 - 1960. New York: American Standards Association.

Andersen, T. and Jensen, K. (2001). Phase modeling of instrument sounds based on psychoacoustic experiments, *Workshop on current research directions in computer music*, Barcelona, Spain.

Ando, S. and Yamaguchi, K. (1993). Statistical study of spectral parameters in musical instrument tones, *Journal of the Acoustical Society of America* **94**(1): 37–45.

Baskent, D. and Shannon, R. (2005). Interactions between cochlear implant electrode insertion depth and frequency-place mapping, *Journal of the Acoustical Society of America* **117**(3 I): 1405–1416.

Beauchamp, J. (1993). Unix workstation software for analysis, graphics, modification, and synthesis of musical sounds, *Audio Engineering Society*, Preprint 3960, pp. 1–17.

Beauchamp, J. and Lakatos, S. (2002). New spectro-temporal measures of musical instrument sounds used for a study of timbral similarity of rise time- and centroid-normalised musical sounds, *Proceedings of the 7th International Conference on Music Perception and Cognition*, Sydney, Australia, pp. 592–595.

Bensa, J., Jensen, K. and Kronland-Martinet, R. (2004). A hybrid resynthesis model for hammer-string interaction of piano tones, *EURASIP Journal of Applied Signal Processing* **7**: 1021–1035.

Bingabr, M., Espinoza-Varas, B. and Loizou, P. (2008). Simulating the effect of spread of excitation in cochlear implants, *Hearing Research* **241**: 73–79.

Blamey, P., Dowell, R., Tong, Y. and Clark, G. (1984). An acoustic model of a multiple-channel cochlear implant, *Journal of the Acoustical Society of America* **76**(1): 97–103.

Bregman, A. S. (2001). *Auditory Scene Analysis: The Perceptual Organization of Sound*, 2nd edn, MIT Press, Massachusetts Institute of Technology.

Bruce, L., White, M., Irlicht, L., O'Leary, S., Dynes, S., Javel, E. and Clark, G. (1999). A stochastic model of the electrically stimulated auditory nerve: Single-pulse response, *IEEE Transactions on Biomedical Engineering* **46**(6): 617–629.

Busby, P., Tong, Y. and Clark, G. (1993). The perception of temporal modulations by cochlear implant patients, *Journal of the Acoustical Society of America* **94**(1): 124–131.

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

Chatterjee, M., Sarampalis, A. and Oba, S. (2006). Auditory stream segregation with cochlear implants: A preliminary report, *Hearing Research* **222**: 100–107.

Clark, G. (2003). *Cochlear Implants: Fundamentals and Applications*, Modern Acoustics and Signal Processing, Springer-Verlag, New York.

Clark, G. M. (1996). Electrical stimulation of the auditory nerve: The coding of frequency, the perception of pitch and the development of cochlear implant speech processing strategies for profoundly deaf people, *Clinical and Experimental Pharmacology and Physiology* **23**: 766–776.

Clarkson, M., Clifton, R. and Perris, E. (1988). Infant timbre perception: Discrimination of spectral envelopes, *Perception and Psychophysics* **43**: 15–20.

Cochlear Pty Ltd (2002). *ACE and CIS DSP strategies. Software requirements specification*. N95287F Issue 1.

Conning, M. (2005). *Acoustic modelling of cochlear implants*, Master's thesis (Bio-Engineering), Faculty of Engineering, Built Environment and Information Technology, University of Pretoria.

Cooper, H. and Roberts, B. (2007). Auditory stream segregation of tone sequences in cochlear implant listeners, *Hearing Research* **225**: 11–24.

Dannenbring, G. and Bregman, A. (1976). Effect of silence between tones on auditory stream segregation, *Journal of the Acoustical Society of America* **59**(4): 987–989.

De Poli, G. and Prandoni, P. (1997). Sonological models for timbre characterization, *Journal of New Music Research* **26**: 170–197.

Donnelly, P. and Limb, C. (2009). Music perception in cochlear implant users, *in* J. Niparko (ed.), *Cochlear Implants: Principles and Practices*, 2nd edn, Lippincott Williams & Wilkins, Philadelphia, pp. 223–228.

Dorman, M., Loizou, P. and Rainey, D. (1997a). Simulating the effect of cochlear-implant electrode insertion depth on speech understanding, *Journal of the Acoustical Society of America* **102**(5): 2993–2996.

Dorman, M., Loizou, P. and Rainey, D. (1997b). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs, *Journal of the Acoustical Society of America* **102**(4): 2403–2411.

Emiroglu, S. and Kollmeier, B. (2008). Timbre discrimination in normal-hearing and hearing-impaired listeners under different noise conditions, *Brain Research* **1220**(C): 199–207.

Faulkner, A., Rosen, S. and Stanton, D. (2003). Simulations of tonotopically mapped speech processors for cochlear implant electrodes varying in insertion depth, *Journal of the Acoustical Society of America* **113**(2): 1073–1080.

Fearn, R. (2001). *Music and pitch perception of cochlear implant recipients*, PhD thesis, School of Physics, University of New South Wales.

Fishman, K., Shannon, R. and Slattery, W. (1997). Speech recognition as a function of the number of electrodes used in the speak cochlear implant speech processor, *Journal of Speech, Language and Hearing Research* **40**: 1201–1215.

Fletcher, N. and Rossing, T. (1998). *The Physics of Musical Instruments*, 2nd edn, Springer-Verlag, New York.

Friesen, L., Shannon, R., Baskent, D. and Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants, *Journal of the Acoustical Society of America* **110**(2): 1150–1163.

Fritts, L. (No date). The University of Iowa Electronic Music Studios Database [Online], Available from http://theremin.music.uiowa.edu/MIS.html. [Accessed: 20 February 2009].

Fu, Q.-J. and Galvin III, J. (2001). Recognition of spectrally asynchronous speech by normal-hearing listeners and nucleus-22 cochlear implant users, *Journal of the Acoustical Society of America* **109**(3): 1166–1172.

Fu, Q.-J. and Shannon, R. (1998). Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners, *Journal of the Acoustical Society of America* **104**(5): 2570–2577.

Fu, Q.-J. and Shannon, R. (1999). Effect of acoustic dynamic range on phoneme recognition in quiet and noise by cochlear implant users, *Journal of the Acoustical Society of America* **106**(6): 65–70.

Galvin III, J., Fu, Q.-J. and Nogaki, G. (2007). Melodic contour identification by cochlear implant listeners, *Ear and Hearing* **28**: 302–319.

Galvin III, J., Fu, Q.-J. and Oba, S. (2008). Effect of instrument timbre on melodic contour identification by cochlear implant users, *Journal of the Acoustical Society of America* **124**(4): 189–195.

Gelfand, S. (1990). Theory of signal detection, *Hearing: An introduction to psychological and physiological acoustics*, Marcel Dekker, Inc., New York, pp. 313–324.

Getty, D., Swets, J. and Swets, J. (1980). The observer's use of perceptual dimensions in signal identification, *in* R. S. Nickerson (ed.), *Attention and Performance VIII*, Lawrence Erlbaum Associates, pp. 361–380.

Getty, D., Swets, J., Swets, J. and Green, D. (1979). On the prediction of confusion matrices from similarity judgements, *Perception and Psychophysics* **26**(1): 1–19.

Gfeller, K. and Lansing, C. (1991). Melodic, rhythmic, and timbral perception of adult cochlear implant users, *Journal of Speech and Hearing Research* **34**: 916–920.

Gfeller, K., Christ, A., Knutson, J., Witt, S. and Mehr, M. (2003). The effects of familiarity and complexity on appraisal of complex songs by cochlear implant recipients and normal hearing adults, *Journal of Music Therapy* **40**(2): 78–112.

Gfeller, K., Christ, A., Knutson, J., Witt, S., Murray, K. and Tyler, R. (2000). Musical backgrounds, listening habits, and aesthetic enjoyment of adult cochlear implant recipients, *Journal of the American Academy of Audiology* **11**: 390–406.

Gfeller, K., Knutson, J., Woodworth, G., Witt, S. and DeBus, B. (1998). Timbral recognition and appraisal by adult cochlear implant users and normal-hearing adults, *Journal of the American Academy of Audiology* **9**(1): 1–19.

Gfeller, K., Olszewski, C., Rychener, M., Sena, K., Knutson, J., Witt, S. and Macpherson, B. (2005). Recognition of "real-world" musical excerpts by cochlear implant recipients and normal-hearing adults, *Ear and Hearing* **26**: 237–250.

Gfeller, K., Turner, C., Mehr, M., Woodworth, G., Fearn, R., Knutson, J., Witt, S. and Stordahl, J. (2002a). Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults, *Cochlear Implants International* **3**(1): 29–53.

Gfeller, K., Turner, C., Oleson, J., Zhang, X., Gantz, B., Froman, R. and Olszewski, C. (2007). Accuracy of cochlear implant recipients on pitch perception, melody recognition, and speech reception in noise, *Ear and Hearing* **28**(3): 412–423.

Gfeller, K., Witt, S., Adamek, M., Mehr, M., Rogers, J., Stordahl, J. and Ringgenberg, S. (2002b). Effects of training on timbre recognition and appraisal by postlingually deafened cochlear implant recipients, *Journal of the American Academy of Audiology* **13**(3): 132–145.

Gfeller, K., Witt, S., Kim, K.-H., Adamek, M. and Coffman, D. (1999). A computerized music training program for cochlear implant recipients, *Journal of the Academy of Rehabilitative Audiology* **32**: 11–27.

Gfeller, K., Witt, S., Woodworth, G., Mehr, M. and Knutson, J. (2002c). Effects of frequency, instrumental family, and cochlear implant type on timbre recognition and appraisal, *Annals of Otology, Rhinology and Laryngology* **111**(4): 349–356.

Gfeller, K., Woodworth, G., Robin, D., Witt, S. and Knutson, J. (1997). Perception of rhythmic and sequential pitch patterns by normally hearing adults and adult cochlear implant users, *Ear and Hearing* **18**: 252–260.

Green, D. and Swets, J. (1966). *Signal detection theory and psychophysics*, John Wiley and Sons, New York.

Greenwood, D. (1990). A cochlear frequency-position function for several species - 29 years later, *Journal of the Acoustical Society of America* **87**(6): 2592–2602.

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

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

Hall, M. and Beauchamp, J. (2009). Clarifying spectral and temporal dimensions of musical instrument timbre, *Canadian Acoustics - Acoustique Canadienne* **37**(1): 3–22.

Hanekom, J. and Hugo, S. (2010). Modelling of timbre perception of cochlear implantees, presented at CI 2010, *the 11th International Conference on Cochlear Implants and other Auditory Implantable Technologies*, Stockholm, Sweden.

Hartmann, W. M. (2005). *Signals, Sound, and Sensation*, Modern Acoustics and Signal Processing, Springer Science and Business Media, New York.

Henry, B. and Turner, C. (2003). The resolution of complex spectral patterns by cochlear implant and normal-hearing listeners, *Journal of the Acoustical Society of America* **113**(5): 2861–2873.

Herrera-Boyer, P., Peeters, G. and Dubnov, S. (2003). Automatic classification of musical instrument sounds, *Journal of New Music Research* **32**(1): 3–21.

Holden, L., Skinner, M., Holden, T. and Demorest, M. (2002). Effects of stimulation rate with the nucleus 24 ace speech coding strategy, *Ear and Hearing* **23**(5): 463–476.

Hopkins, K. and Moore, B. (2007). Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information, *Journal of the Acoustical Society of America* **122**(2): 1055–1068.

Horner, A., Beauchamp, J. and So, R. (2009). Detection of time-varying harmonic amplitude alterations due to spectral interpolations between musical instrument tones, *Journal of the Acoustical Society of America* **125**(1): 492–502.

House, W. and Berliner, K. (1982). Cochlear implants: progress and perspectives, *Annals of Otology, Rhinology and Laryngology Supplement* **91**: 1–124.

Houtsma, A. (1997). Pitch and timbre: definition, meaning and use, *Journal of New Music Research* **26**: 104–115.

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

Jensen, K. (1996). The control of musical instruments, *Nordic Acoustical Meeting*, Helsinki.

Jensen, K. (1999a). Envelope model of isolated musical sounds, *Proceedings of 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99)*, NTNU, Trondheim.

Jensen, K. (1999b). *Timbre models of musical sounds: from the model of one sound to the model of one instrument*, PhD thesis, Department of Computer Science, University of Copenhagen.

Jensen, K. (2001). The timbre model, *Workshop on current research directions in computer music*, Barcelona, Spain.

Jensen, K. (2002a). Musical instruments parametric evolution, *Proceedings of the International Symposium on Musical Acoustics*, Computer Music Association, Mexico City, Mexico, pp. 319–326.

Jensen, K. (2002b). Perceptual and physical aspects of musical sounds, *Journal of Sangeet Research Academy, India* **1**: 1–22.

Jensen, K. and Marentakis, G. (2001). Hybrid perception, *Papers from the 1st Seminar on Auditory Models*.

Jesteadt, W. (1980). An adaptive procedure for subjective judgements, *Perception and Psychophysics* **28**(1): 85–88.

Kameoka, A. and Kuriyagawa, M. (1969). Consonance theory part II: consonance of complex tones and its calculation method, *Journal of the Acoustical Society of America* **45**(6): 1460–1469.

Koelsch, S. and Siebel, W. (2005). Towards a neural basis of music perception, *Trends in Cognitive Sciences* **9**(12): 578–584.

Koelsch, S., Wittfoth, M., Wolf, A., Müller, J. and Hahne, A. (2004). Music perception in cochlear implant users: an event-related potential study, *Clinical Neurophysiology* **115**: 966–972.

Kong, Y., Cruz, R., Jones, J. and Zeng, F. (2004). Music perception with temporal cues in acoustic and electric hearing, *Ear and Hearing* **25**: 173–185.

Krimphoff, J., McAdams, S. and Winsberg, S. (1994). Caractérisation du timbre des sons complexes. II. Analyses acoustiques et quantification psychophysique, *Journal de Physique* **4**(C5): 625–628.

Krumhansl, C. (1989). Why is musical timbre so hard to understand?, *Structure and Perception of Electroacoustic Sound and Music: Proceedings of the Marcus Wallenberg Symposium Held in Lund*, pp. 43–53.

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

Laneau, J. and Wouters, J. (2004). Relative contributions of temporal and place pitch cues to fundamental frequency discrimination in cochlear implantees, *Journal of the Acoustical Society of America* **116**(6): 3606–3619.

Lassaletta, L., Castro, A., Bastarrica, M., Pérez-Mora, R., Madero, R., de Sarriá, J. and Gavilán, J. (2007). Does music perception have an impact on quality of life following implantation?, *Acta Oto-Laryngologica* **127**: 682–686.

Leal, M., Shin, Y., Laborde, M., Calmels, M., Verges, S., Lugardon, S., Andrieu, S., Deguine, O. and Fraysse, B. (2003). Music perception in adult cochlear implant recipients, *Acta Oto-Laryngologica* **123**: 826–835.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics, *Journal of the Acoustical Society of America* **49**(2): 467–477.

Limb, C. (2006). Cochlear implant-mediated perception of music, *Current Opinion in Otolaryngology and Head and Neck Surgery* **14**: 337–340.

Lindeberg, T. (1996). Edge detection and ridge detection with automatic scale selection, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **1**: 465–470.

Lister, J., Koehnke, J. and Besing, J. (2000). Binaural gap duration discrimination in listeners with impaired hearing and normal hearing, *Ear and Hearing* **21**(2): 141–150.

Loizou, P. (1998). Mimicking the human ear, *IEEE Signal Processing Magazine* **15**(5): 101–130.

Loizou, P. (1999a). Introduction to cochlear implants, *IEEE Engineering in Medicine and Biology Magazine* **18**(1): 32–42.

Loizou, P. and Poroy, O. (2001). Minimum spectral contrast needed for vowel identification by normal-hearing and cochlear implant listeners, *Journal of the Acoustical Society of America* **110**(3): 1619–1627.

Loizou, P. C. (1999b). Signal-processing techniques for cochlear implants, *IEEE Engineering in Medicine and Biology* **18**(3): 34–46.

Loizou, P., Dorman, M. and Fitzke, J. (2000a). The effect of reduced dynamic range on speech understanding: Implications for patients with cochlear implants, *Ear and Hearing* **21**(1): 25–31.

Loizou, P., Poroy, O. and Dorman, M. (2000b). The effect of parametric variations of cochlear implant processors on speech understanding, *Journal of the Acoustical Society of America* **108**(2): 790–802.

Loureiro, M., de Paula, H. and Yehia, H. (2004). Timbre classification of a single musical instrument, *Proceedings of the International Conference on Music Information Retrieval*.

Marentakis, G. and Jensen, K. (2001). The timbre engine - progress report, *Workshop on current research directions in computer music*, Barcelona, Spain.

Martin, K. D., Scheirer, E. D. and Vercoe, B. L. (1998). Music content analysis through models of audition, *ACM Multimedia '98 Workshop on Content Processing of Music for Multimedia Applications*, Bristol, U.K.

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

McAulay, R. and Quatieri, T. (1986). Speech analysis/synthesis based on a sinusoidal representation, *IEEE Transactions on Acoustics, Speech and Signal Processing* **34**(4): 744–754.

McDermott, H. (2004). Music perception with cochlear implants: a review, *Trends in Amplification* **8**(2): 49–82.

McDermott, H. and Looi, V. (2004). Perception of complex signals, including musical sounds, with cochlear implants, *International Congress Series* **1273**: 201–204.

McDermott, H. and McKay, C. (1997). Musical pitch perception with electrical stimulation of the cochlea, *Journal of the Acoustical Society of America* **101**(3): 1622–1631.

McKay, C. (2005). Spectral processing in cochlear implants, *International Review of Neurobiology* **70**: 473–509.

McKay, C. and McDermott, H. (1996). The perception of temporal patterns for electrical stimulation presented at one or two intracochlear sites, *Journal of the Acoustical Society of America* **100**(2 I): 1081–1092.

Miller, G. and Nicely, P. (1955). An analysis of perceptual confusions among some English consonants, *Journal of the Acoustical Society of America* **27**(2): 338–352.

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

Moore, B., Glasberg, B. and Hopkins, K. (2006). Frequency discrimination of complex tones by hearing-impaired subjects: Evidence for loss of ability to use temporal fine structure, *Hearing Research* **222**(1-2): 16–27.

Nie, K., Barco, A. and Zeng, F. (2006). Spectral and temporal cues in cochlear implant speech perception, *Ear and Hearing* **27**: 208–217.

Nimmons, G., Kang, R., Drennan, W., Longnion, J., Ruffin, C., Worman, T., Yueh, B. and Rubinstein, J. (2008). Clinical assessment of music perception in cochlear implant listeners, *Otology and Neurotology* **29**: 149–155.

Nogueira, W., Büchner, A., Lenarz, T. and Edler, B. (2005). A psychoacoustic "NofM"-type speech coding strategy for cochlear implants, *EURASIP Journal of Applied Signal Processing* **18**: 3044–3059.

Patel, A. (2003). Language, music, syntax and the brain, *Nature Neuroscience* **6**(7): 674–681.

Peretz, I. and Coltheart, M. (2003). Modularity of music processing, *Nature Neuroscience* **6**(7): 688–691.

Pijl, S. and Schwarz, D. (1995). Melody recognition and musical interval perception by deaf subjects stimulated with electrical pulse trains through single cochlear implant electrodes, *Journal of the Acoustical Society of America* **98**: 886–895.

Plomp, R. (1969). Timbre as a multidimensional attribute of complex tones, *in* R. Plomp and G. Smoorenburg (eds), *The Proceedings of the International Symposium on Frequency Analysis and Periodicity Detection in Hearing*, Driebergen, The Netherlands, pp. 397–414.

Pressnitzer, D., Bestel, J. and Fraysse, B. (2005). Music to electric ears: pitch and timbre perception by cochlear implant patients, *Annals New York Academy of Sciences* **1060**: 343–345.

Pretorius, L. and Hanekom, J. (2005). Perception of speech and music sounds: implications for cochlear implants, Unpublished article.

Rasch, R. and Plomp, R. (1982). The perception of musical tones, *in* D. Deutsch (ed.), *The Psychology of Music*, Academic Press, chapter 1, pp. 1–24.

Risset, J.-C. and Wessel, D. L. (1982). Exploration of timbre by analysis and synthesis, *in* D. Deutsch (ed.), *The Psychology of Music*, Academic Press, pp. 25–58.

Ross, M. (2008). Listening to music through a cochlear implant (part 1), *Hearing Loss Magazine*, pp. 21–23.

Rubinstein, J. and Turner, C. (2003). A novel acoustic simulation of cochlear implant hearing: effects of temporal fine structure, *Proceedings of the 1st International IEEE EMBS Conference, Conference on Neural Engineering*, Capri Island, Italy, pp. 142–145.

Samson, S., Zattore, R. and Ramsay, J. (1997). Multidimensional scaling of synthetic musical timbre: perception of spectral and temporal characteristics, *Canadian Journal of Experimental Psychology* **51**(4): 307–315.

Shannon, R. (1983). Multi-channel electrical stimulation of the auditory nerve in man. I. Basic psychophysics, *Hearing Research* **11**: 157–189.

Shannon, R. (1989). Detection of gaps in sinusoids and pulse trains by patients with cochlear implants, *Journal of the Acoustical Society of America* **85**(6): 2587–2592.

Shannon, R. (2005). Speech and music have different requirements for spectral resolution, *International Review of Neurobiology* **70**: 121–134.

Shannon, R., Adams, D., Ferrel, R., Palumbo, R. and Grandgenett, M. (1990). A computer interface for psychophysical and speech research with the nucleus cochlear implant, *Journal of the Acoustical Society of America: Technical Notes and Research Briefs* **87**(2): 905–907.

Singh, P. and Bregman, A. (1997). The influence of different timbre attributes on the perceptual segregation of complex-tone sequences, *Journal of the Acoustical Society of America* **102**(4): 1943–1952.

Skinner, M., Holden, L., Whitford, L., Plant, K., Psarros, C. and Holden, T. (2002). Speech recognition with the nucleus 24 speak, ace, and cis speech coding strategies in newly implanted adults, *Ear and Hearing* **23**: 207–223.

Stainsby, T., McDermott, H., McKay, C. and Clark, G. (2002). Musical timbre perception with cochlear implants: investigations using forward masking, *Proceedings of the 7th International Conference on Music Perception and Cognition*, Sydney, Australia, pp. 604–607.

Stewart, J. (1999). *Calculus: Early Transcendentals*, 4th edn, Brooks/Cole Publishing Company, Pacific Grove, California.

Strong, W. and Plitnik, G. (1992). *Music, speech, audio*, Soundprint, Provo, Utah.

Svirsky, M. (2000). Mathematical modeling of vowel perception by users of analog multichannel cochlear implants: Temporal and channel-amplitude cues, *Journal of the Acoustical Society of America* **107**(3): 1521–1529.

Terasawa, H., Slaney, M. and Berger, J. (2005). The thirteen colors of timbre, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, pp. 323–326.

Terasawa, H., Slaney, M. and Berger, J. (2006). A statistical model of timbre perception, *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*.

Throckmorton, C. and Collins, L. (2002). The effect of channel interactions on speech recognition in cochlear implant subjects: Predictions from an acoustic model, *Journal of the Acoustical Society of America* **112**(1): 285–296.

Turner, C. and Holte, L. (1987). Discrimination of spectral-peak amplitude by normal and hearing-impaired subjects, *Journal of the Acoustical Society of America* **81**(2): 445–451.

Van Hoesel, R. and Tyler, R. (2003). Speech perception, localization, and lateralization with bilateral cochlear implants, *Journal of the Acoustical Society of America* **113**(3): 1617–1630.

van Tasell, D., Soli, S., Kirby, V. and Widin, G. (1987). Speech waveform envelope cues for consonant recognition, *Journal of the Acoustical Society of America* **82**(4): 1152–1161.

van Wieringen, A. and Wouters, J. (1999). Natural vowel and consonant recognition by laura cochlear implantees, *Ear and Hearing* **20**(2): 89–103.

van Zyl, J. (2008). *Objective determination of vowel intelligibility of a cochlear implant model*, Master's thesis (Bio-Engineering), Faculty of Engineering, Built Environment and Information Technology, University of Pretoria.

Vandali, A., Whitford, L., Plant, K. and Clark, G. (2000). Speech perception as a function of electrical stimulation rate: Using the nucleus 24 cochlear implant system, *Ear and Hearing* **21**(6): 608–624.

Vanpoucke, F., Zarowski, A. and Peeters, S. (2004). Identification of the impedance model of an implanted cochlear prosthesis from intracochlear potential measurements, *IEEE Transactions on Biomedical Engineering* **51**(12): 2174–2183.

Wang, M. and Bilger, R. (1973). Consonant confusions in noise: a study of perceptual features, *Journal of the Acoustical Society of America* **54**(5): 1248–1266.

Wedin, L. and Goude, G. (1972). Dimension analysis of the perception of instrumental timbre, *Scandinavian Journal of Psychology* **13**: 228–240.

White, M., Merzenich, M. and Gardi, J. (1984). Multichannel cochlear implants: channel interactions and processor design, *Archives of Otolaryngology* **110**: 493–501.

Whitford, L., Seligman, P., Blamey, P., McDermott, H. and Patrick, J. (1993). Comparison of current speech coding strategies, *Advances in oto-rhino-laryngology* **48**: 85–90.

Whitford, L., Seligman, P., Everingham, C., Antognelli, T., Skok, M., Hollow, R., Plant, K., Gerin, E., Staller, S., McDermott, H., Gibson, W. and Clark, G. (1995). Evaluation of the nucleus spectra 22 processor and new speech processing strategy (speak) in postlinguistically deafened adults, *Acta Oto-Laryngologica* **115**: 629–637.

Willems, P. (2000). Genetic causes of hearing loss, *The New England Journal of Medicine* **342**: 1101–1109.

Wilson, B. (2006). Speech processing strategies, *in* H. Cooper and L. Craddock (eds), *Cochlear Implants: A Practical Guide*, 2nd edn, Whurr Publishers Ltd., London and Philadelphia.

Wilson, B. and Dorman, M. (2008). Cochlear implants: a remarkable past and a brilliant future, *Hearing Research* **242**: 3–21.

Zattore, R. (2001). Neural specializations for tonal processing, *Annals of the New York Academy of Sciences* **930**(1): 193–210.

Zattore, R., Belin, P. and Penhune, V. (2002). Structure and function of auditory cortex: music and speech, *Trends in Cognitive Sciences* **6**(1): 37–46.

Zeng, F.-G., Grant, G., Niparko, J., Galvin, J., Shannon, R., Opie, J. and Segel, P. (2002). Speech dynamic range and its effect on cochlear implant performance, *Journal of the Acoustical Society of America* **111**(1): 377–386.

Zwicker, E. and Fastl, H. (1999). *Psychoacoustics: Facts and Models*, Springer series in information sciences, 2nd edn, Springer-Verlag.

APPENDIX A

ADDITIONAL MUSICAL INSTRUMENT SOUNDS

Illustrations of the 6 musical instrument sounds that were used in this study in addition to the four sounds used as examples throughout the dissertation are presented in this Appendix. The four primary sounds presented in the body of this dissertation are the piano, trumpet, clarinet and violin. The additional 6 sounds illustrated here are the French horn, trombone, flute saxophone, cello and viola. For each sound, the time domain, frequency domain and additive parameter representations are given for both the original sounds and the sounds processed through the acoustic model.

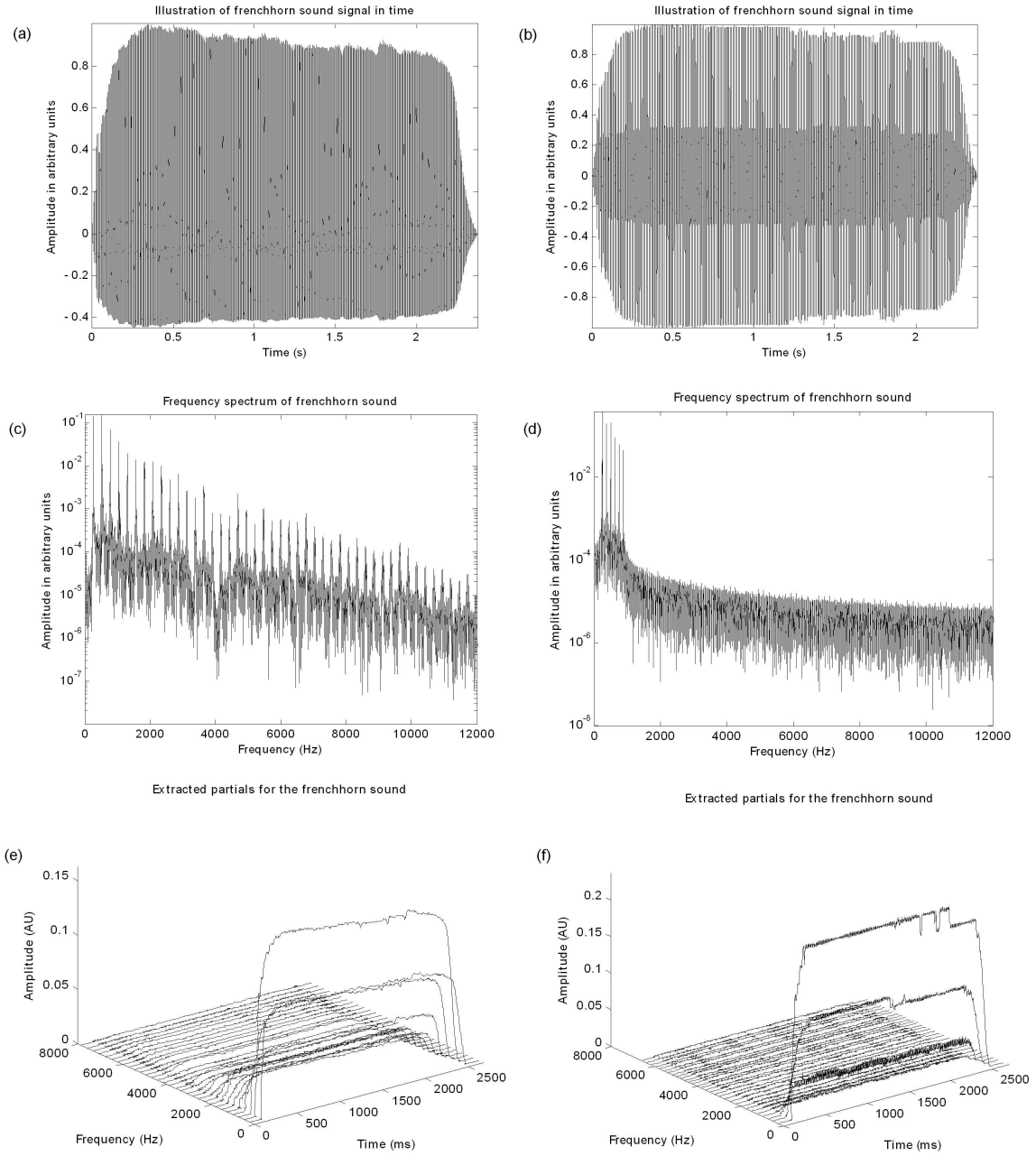

Figure A.1.
Illustrations of the French horn sound, showing the time domain representation of (a) the original sound and (b) the processed sound, the frequency domain representation of (c) the original sound and (d) the processed sound, and the additive parameter representations of (e) the original sound and (f) the processed sound.

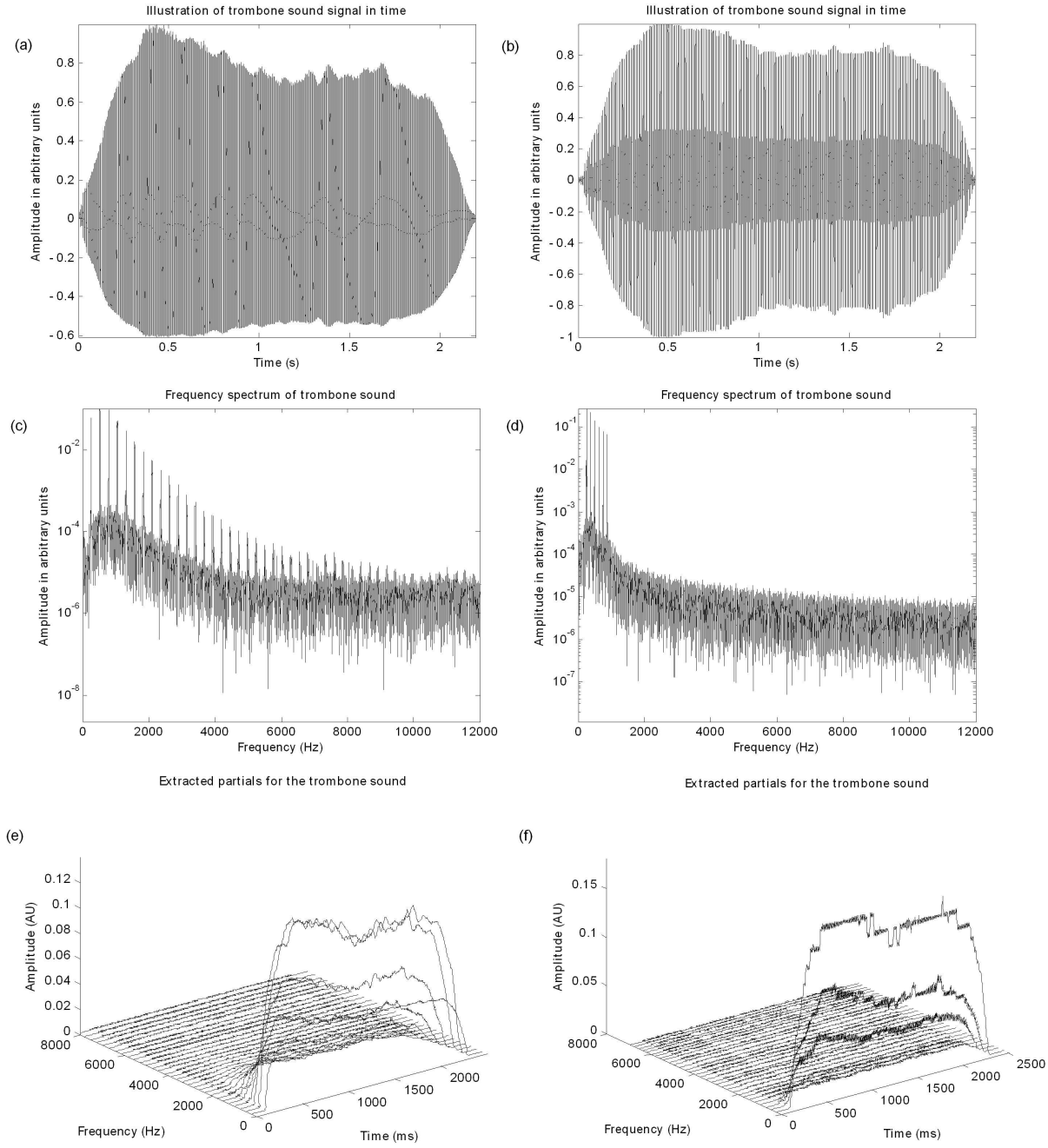

Figure A.2.
Illustrations of the trombone sound, showing the time domain representation of (a) the original sound and (b) the processed sound, the frequency domain representation of (c) the original sound and (d) the processed sound, and the additive parameter representations of (e) the original sound and (f) the processed sound.

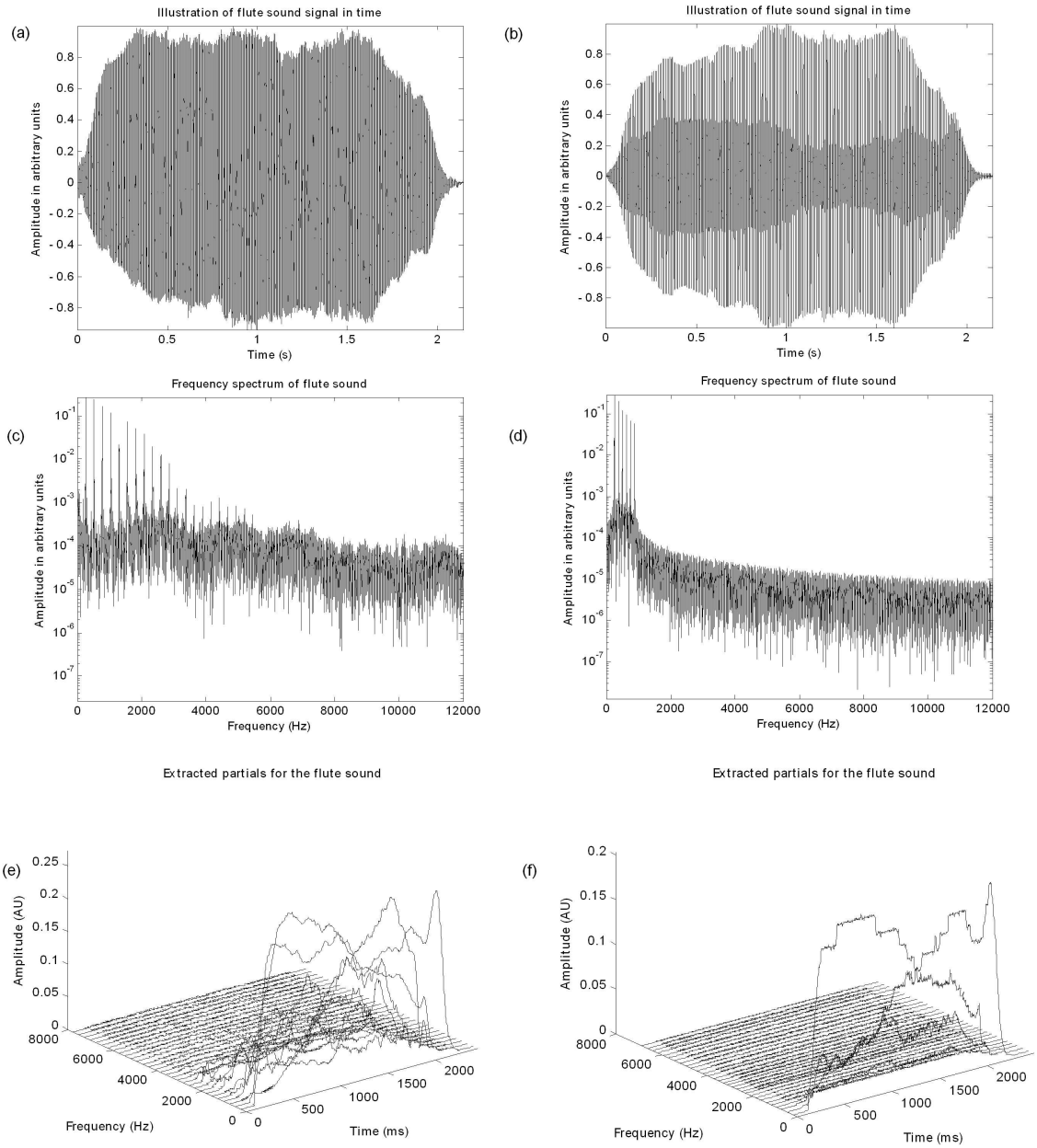

Figure A.3.
Illustrations of the flute sound, showing the time domain representation of (a) the original sound and (b) the processed sound, the frequency domain representation of (c) the original sound and (d) the processed sound, and the additive parameter representations of (e) the original sound and (f) the processed sound.

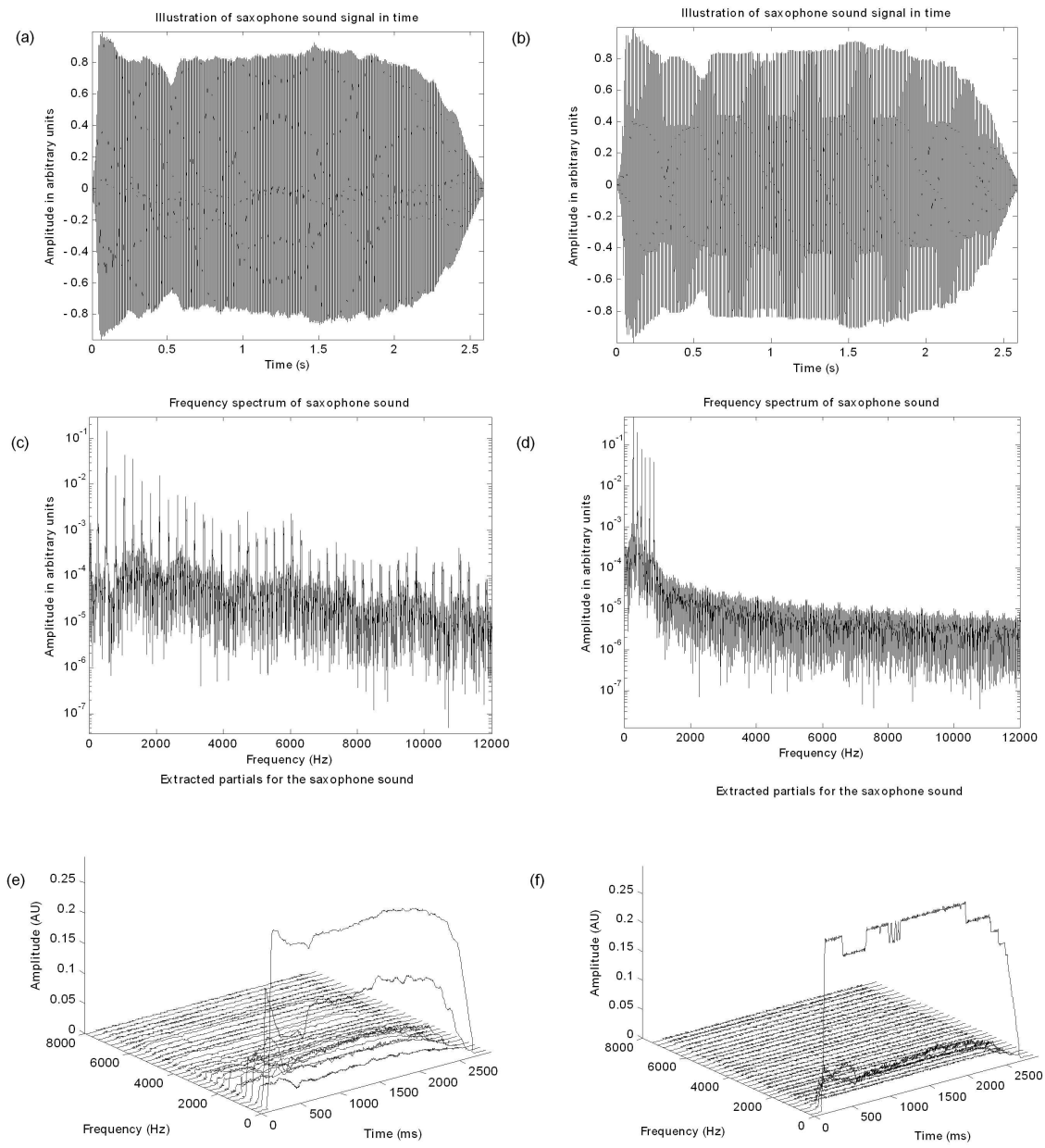

Figure A.4.
Illustrations of the saxophone sound, showing the time domain representation of (a) the original sound and (b) the processed sound, the frequency domain representation of (c) the original sound and (d) the processed sound, and the additive parameter representations of (e) the original sound and (f) the processed sound.

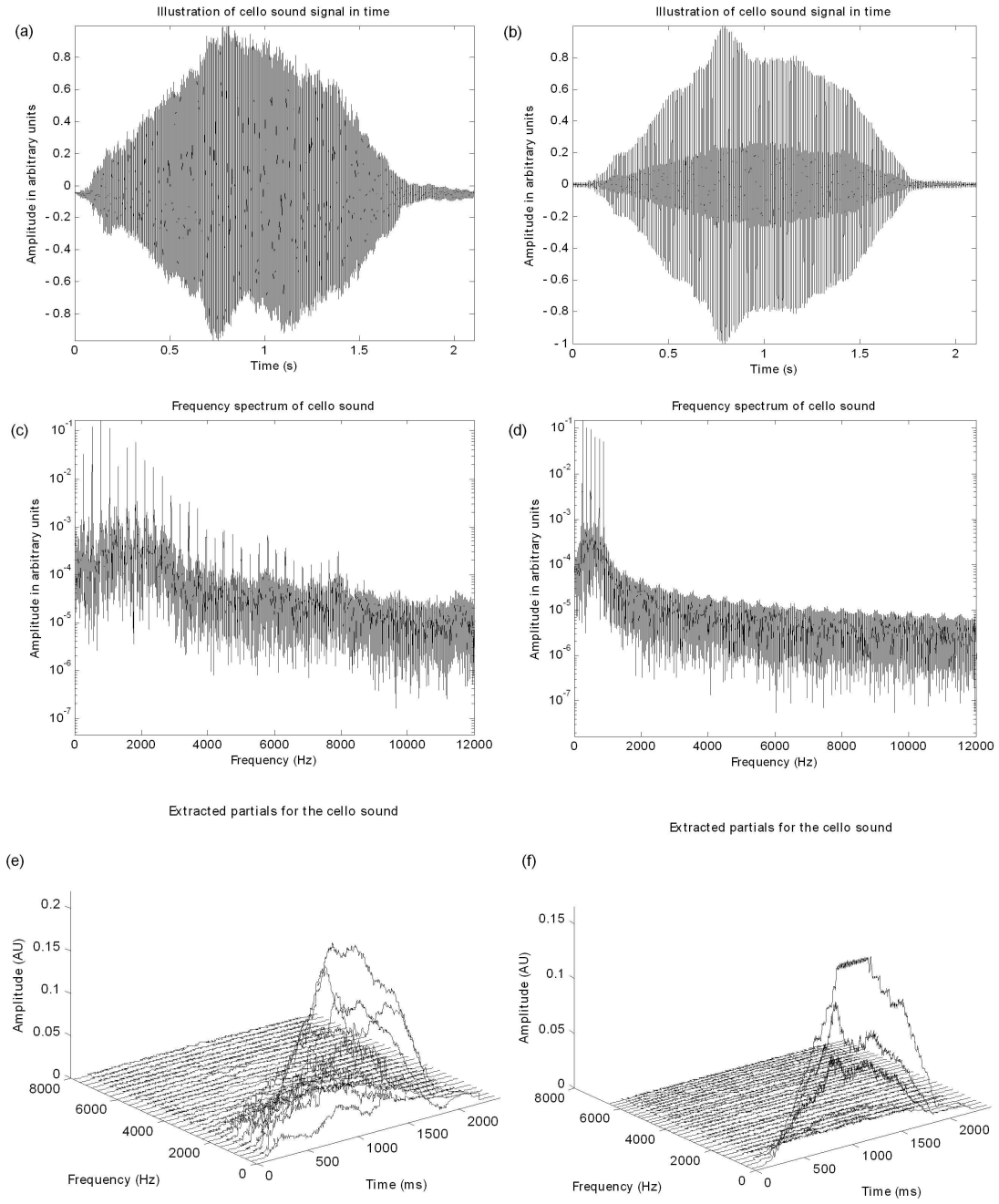

Figure A.5.
Illustrations of the cello sound, showing the time domain representation of (a) the original sound and (b) the processed sound, the frequency domain representation of (c) the original sound and (d) the processed sound, and the additive parameter representations of (e) the original sound and (f) the processed sound.

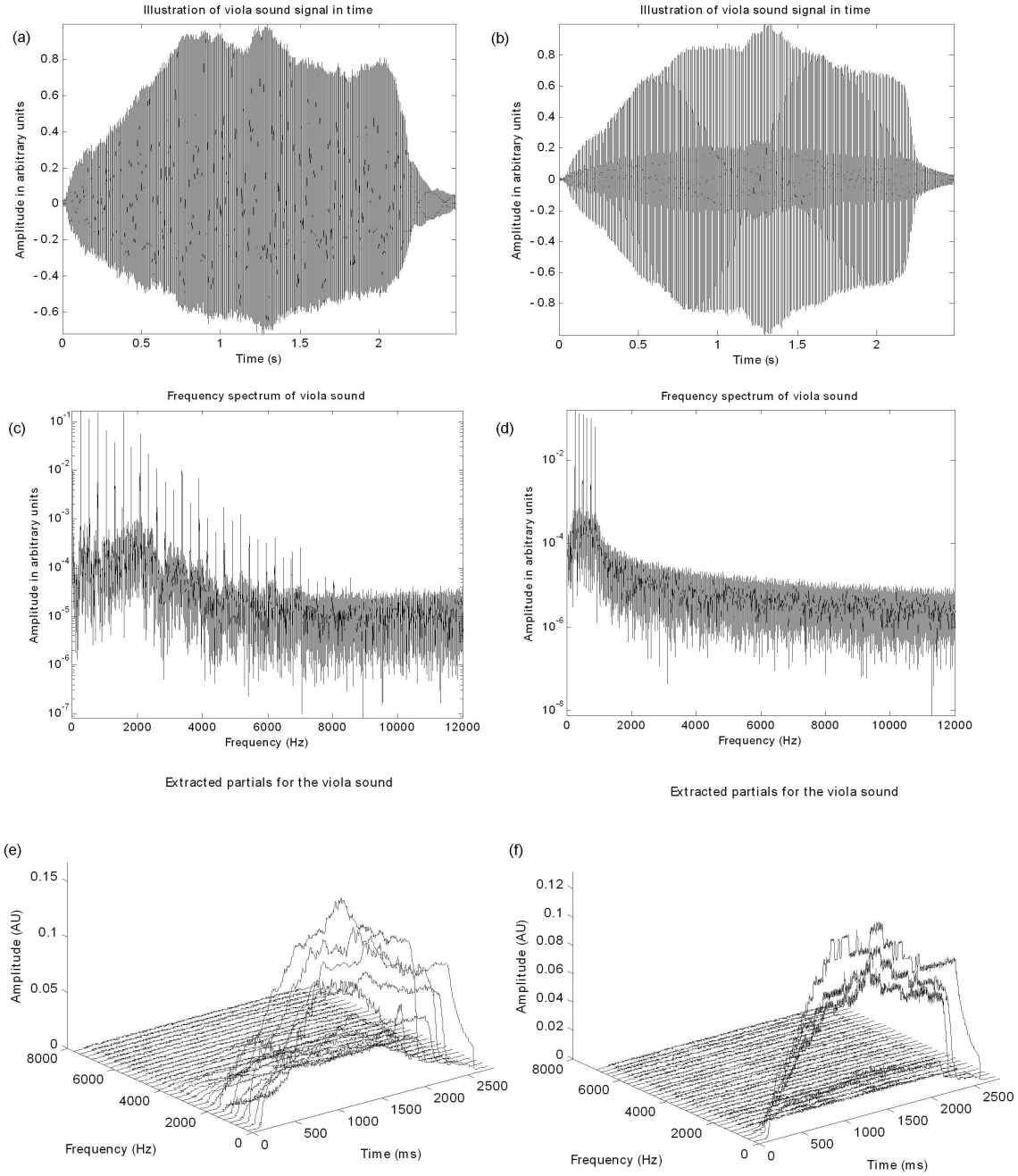

Figure A.6.
Illustrations of the viola sound, showing the time domain representation of (a) the original sound and (b) the processed sound, the frequency domain representation of (c) the original sound and (d) the processed sound, and the additive parameter representations of (e) the original sound and (f) the processed sound.